Estadística Descriptiva y Probabilidad

(Teoría y problemas)

3ª Edición

Autores
I. Espejo Miranda
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
A. M. Rodríguez Chía
A. Sánchez Navas
C. Valero Franco

UCA Universidad de Cádiz

Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN: 978-84-9828-058-6

Depósito legal:

Parte A

Estadística Descriptiva

Estadística Descriptiva y Probabilidad. Teoría y Problemas (Revisión: Febrero 2006)
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C Valero Franco

Introducción a la estadística descriptiva

La primera parte de este libro está dedicada a la estadística descriptiva. Atendiendo a lo que tradicionalmente se ha entendido por descriptiva se estaría hablando de un conjunto de herramientas, formado por coeficientes y técnicas, que tratan de resumir la información contenida en un conjunto de datos. Sin embargo, la estadística descriptiva es mucho más que eso, en realidad es una parte fundamental de cualquier análisis estadístico complejo, en la que se empiezan a tomar decisiones que afectarán al conjunto de la investigación. Los coeficientes descriptivos darán información sobre la estructura de la población que se estudia, indicando, por ejemplo, si ésta es simétrica, si realmente se trata de una única población o hay una superposición de poblaciones, también pueden detectarse valores extraordinariamente raros, etc.

Desde otra óptica, la mayoría de los coeficientes descriptivos tendrán su homólogo inferencial o poblacional, que necesariamente deberán ser estudiados a la luz de aquellos. Haciendo una pequeña abstracción muchos de los coeficientes descriptivos, los más importantes, se convierten en poblacionales al sustituir frecuencias por probabilidades.

En resumen, el análisis descriptivo es una parte inseparable de cualquier análisis estadístico, que puede tener su continuidad con un análisis inferencial cuando los datos que se manejan se corresponden con una muestra probabilística extraída de una población.

Esta primera parte del libro está compuesta por tres capítulos, el

primero de ellos aborda el problema unidimensional. Se trata de identificar la información que se va a analizar, bien sean variables cuantitativas o de clase, procediéndose a organizarla en distribuciones de frecuencias. Se indica que la primera toma de contacto con las peculiaridades de una distribución se obtiene a través de sus representaciones gráficas y se da al menos una representación para cada uno de los tipos de datos que se manejan. Se calculan todos los coeficientes tradicionales: medidas de centralización, de posición, de dispersión y de forma; se obtienen los momentos respecto al origen y respecto a la media, indicándose que generalizan la mayoría de las medidas anteriores. Se introduce la desigualdad de Tchebychev, poniéndose de manifiesto la relación existente entre la varianza y la media aritmética. Se estudian las transformaciones de variables, haciendo ver que el objetivo es conseguir distribuciones más regulares, que sean comparables, más simétricas; entre todas las transformaciones se dedica especial atención a la normalización o tipificación. Por último, se hace una breve incursión en el análisis exploratorio de datos, recurriendo a representaciones, como los diagramas de cajas, que resaltan las regularidades y las especificidades del conjunto de datos, entre las que cabe destacar la presencia de observaciones candidatas a ser valores extraños o anómalos.

El capítulo segundo supone una generalización al caso de que conjuntamente se tenga más de una variable, viéndose con detenimiento el caso bivariable y destacando el hecho de la posible existencia de relaciones entre dichas variables. La existencia de dependencias merece una especial atención por las consecuencias que de ella se derivan en muchas técnicas estadísticas. Se introducen coeficientes que expresarán el grado de relación entre las variables, distinguiendo los casos en que éstas sean continuas, ordenadas o de clase; lo que conduce a definir medidas de correlación, concordancia y contingencia o asociación.

En el último capítulo de esta parte se aborda el problema del ajuste y la regresión en el plano, lo que supone un primer acercamiento a la modelización estadística. El desarrollo del tema se hace planteando un modelo lineal, empleándose para la estimación de los parámetros el método de los mínimos cuadrados. El análisis de la bondad del ajuste se realiza a través del coeficiente de determinación. El método de la

regresión a la media permite calibrar la calidad de los posibles ajustes a realizar. También se analizan algunas extensiones a los casos de modelos linealizables y polinomiales.

Estadística Descriptiva y Probabilidad. Teoría y Problemas (Revisión: Febrero 2006)
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C Valero Franco

Capítulo 1

Síntesis de la información

1. Reseña histórica

1.1. Introducción

Al acercarse a una ciencia es interesante indagar en sus raíces históricas para obtener una visión de su naturaleza y de sus objetivos como disciplina científica. El estudio de dichas raíces permitirá entender el grado de desarrollo actual, la relación entre sus distintas partes, comprender su terminología -dado que el nombre de un coeficiente, de una técnica, ... suele estar asociado a su origen histórico-, e incluso prever en que dirección evolucionará. En el caso de la Estadística este estudio retrospectivo es particularmente rico en enseñanzas.

A lo largo de los tiempos han sido muchas las concepciones que se le ha dado a la ciencia Estadística, desde la que la ha entendido como un conjunto de técnicas aplicables a una serie de datos, hasta la que la ha concebido como un proceso de extrapolación de conclusiones de la muestra a la población. Actualmente, no puede entenderse la Estadística como un conjunto de conceptos y expresiones matemáticas abstractas, olvidando las motivaciones históricas sobre las que se construyó y su actual papel esencial en cualquier tipo de investigación empírica, tal y como destaca Kruskal en su Enciclopedia Internacional de Estadística.

1.2. Orígenes de la estadística descriptiva

Los orígenes históricos de la Estadística (descriptiva) hay que buscarlos en los procesos de recogida de datos, censos y registros sistemáticos, asumiendo un papel asimilable a una aritmética estatal para asistir al gobernante, que necesitaba conocer la riqueza y el número de sus súbditos con fines tributarios y políticos.

Los primeros registros de riqueza y población que se conocen se deben a los egipcios. Ramsés II en el 1400 a.C. realizó el primer censo conocido de las tierras de Egipto, no siendo éste, se supone, ni el primero ni el último que se hiciera en las tierras bañadas por el Nilo.

Posteriormente, desde el siglo III a.C., en las civilizaciones china y romana se llevan a cabo censos e inventarios de posesiones, que pueden considerarse precedentes institucionalizados de la recogida de datos demográficos y económicos de los Estados Modernos.

Hay que realizar una mención especial del período helénico, en el que las escuelas matemáticas se suceden. Centros como el de Quios, donde estudió Hipócrates (Hipócrates de Quios) el matemático, considerado como el inventor del método matemático y escuelas como las de Cirene, Megara y al final Atenas, donde se reunen los matemáticos, unos alrededor de Protágoras y otros en torno a Sócrates.

En la Edad Media se vuelve a la utilización de la Aritmética para la recogida de datos, existiendo menos interés por la elucubración matemática abstracta. Es en este período de tiempo cuando Carlomagno ordenó en su "Capitulare de villis" la creación de un registro de todos sus dominios y bienes privados.

En el siglo XVII se producen avances sustanciales, y así, en las universidades alemanas se imparten enseñanzas de "Aritmética Política", término con el que se designa la descripción numérica de hechos de interés para la Administración Pública. Destacados autores de Aritmética Política fueron los ingleses Graunt (1620-1674) y Petty (1623-1687).

Con métodos de estimación en los que cabía la conjetura, la experimentación y la deducción, Graunt llega a estimar tasas de mortalidad para la población londinense, analizando además la verosimilitud de la información de que disponía. Por su parte, Petty, cuyas aportaciones estadísticas fueron menos relevantes, tiene el mérito -en opinión de Gutiérrez Cabria- de proponer la creación de un departamento de estadística, en el que se reuniese información no sólo de carácter demográfico, sino también sobre recaudación de impuestos, educación y comercio. Surge en esta época la conciencia de la necesidad de disponer de información, conciencia que va tomando cuerpo a partir de la segunda mitad del siglo XVII en la mayor parte de las potencias europeas y americanas, considerándose como primera oficina de estadística la instituida en Suecia en 1756.

En España, el interés por las investigaciones estatales nació con la preocupación de los Reyes Católicos por mejorar el estado de las "Cosas Públicas", estableciéndose el primer censo del que se tiene referencia en 1482, elaborado por Alonso de Quintanilla. Durante el siglo XVIII se elaboraron censos como el de Ensenada en 1749 y el de Floridablanca en 1787, con una metodología con visos de modernidad. Los actuales censos de periodicidad decenal empezaron a elaborarse en 1860 a cargo de la Junta General de Estadística.

2. La organización de la información

Los datos constituyen la materia prima de la Estadística, pudiéndose establecer distintas clasificaciones en función de la forma en que éstos vengan dados. Se obtienen datos al realizar cualquier tipo de prueba, experimento, valoración, medición, observación,...

Este capítulo tiene por finalidad la descripción de un conjunto de datos, sin considerar que éstos puedan pertenecer a un colectivo más amplio y, por supuesto, sin la intención de proyectar los resultados que se obtengan al colectivo global; objeto esto último de lo que se conoce como Inferencia Estadística.

2.1. Variable y atributo

Se realiza una primera clasificación del tipo de datos en función de que las observaciones resultantes del experimento sean de tipo cualitativo o cuantitativo, en el primero de los casos se tiene un *atributo* y en el segundo una *variable*. Para hacer referencia genéricamente a una variable o a un atributo se utilizará el término *carácter*.

Ejemplo 1.1 *Como ejemplos de atributos pueden considerarse el color del pelo de un colectivo de personas, su raza o el idioma que hablan y como variables su estatura, peso o edad.*

Para poder operar con un atributo es necesario asignar a cada una de sus clases un valor numérico, con lo que se transforma en una variable, esta asignación se hará de forma que los resultados que se obtengan al final del estudio sean fácilmente interpretables.

Ejercicio 1.1 *Clasifique los siguientes datos según sean variables o atributos:*

a) *El color de ojos de un grupo de 20 personas.*

b) *La nacionalidad de un conjunto de individuos.*

c) *Las dioptrías de un grupo de personas miopes.*

d) *Los matices de color de un cuadro impresionista.*

e) *Las dianas que consigue un arquero sobre un total de 100 intentos.*

2.2. Variables discretas y continuas

Dentro del conjunto de las variables se distingue entre *discretas* y *continuas*. Se dice que una variable es discreta cuando entre dos valores consecutivos no toma valores intermedios y que es continua cuando puede tomar cualquier valor dentro de un intervalo.

Ejemplo 1.2 *La estatura de un grupo de personas sería una variable continua, mientras que el número de cabellos que tienen en la cabeza sería una variable discreta.*

En la práctica todas las variables son discretas debido a la limitación de los aparatos de medida, y así, en el ejemplo de las estaturas, quizás se podría detectar una diferencia de una cienmilésima de metro, o a lo más, de una millonésima, pero dados dos individuos que se diferencien en una millonésima no puede detectarse otro que tenga una estatura intermedia. De todas formas, en general se trata a las variables "teóricamente" continuas como tales, por razones que se pondrán de manifiesto más adelante.

Ejercicio 1.2 *Indique cuáles de las siguientes variables son continuas y cuáles discretas:*

a) *El número de moléculas de agua de un pantano.*

b) *La edad exacta de un grupo de 50 niños.*

c) *La distancia por carretera entre las capitales de provincia peninsulares españolas.*

d) *La distancia al centro de la diana de las flechas lanzadas por un arquero.*

e) *El número de docenas de huevos que se recolecta al día en una granja de gallinas.*

Si la ocasión lo requiere se tiene la posibilidad de transformar una variable discreta en continua o viceversa. Para transformar una variable discreta en continua, una vez ordenados los valores, se asigna a cada uno de ellos un intervalo que tenga por extremos el punto medio respecto al valor anterior y el punto medio respecto al valor siguiente. Esta operación tiene interés, por ejemplo, en la aproximación de distribuciones discretas a continuas, como se tendrá la oportunidad de comprobar en la segunda parte de este manual.

Para transformar una variable continua en discreta basta con hacer corresponder a cada uno de los intervalos su punto medio o *marca de clase*.

Ejercicio 1.3 *Transforme la variable continua que toma valores en los intervalos* $(0, 2]$, $(2, 3]$, $(3, 6]$, $(6, 10]$, $(10, 15]$ *en variable discreta.*

2.3. Clasificación de las series estadísticas

Además de por su naturaleza, se pueden realizar distintas clasificaciones del conjunto de los datos o serie estadística.

1. **Por su número**
   - *a*) Finitas. Las que tienen un número finito de elementos.
   - *b*) Infinitas. Cuando tienen infinitos elementos.
2. **Por su obtención**
   - *a*) Objetivas. Obtenidas con métodos exactos de medición.
   - *b*) Subjetivas. Obtenidas mediante apreciaciones personales.
3. **Por su dimensión**
   - *a*) Unidimensionales: $x_1, x_2, x_3, \cdots, x_n$.
   - *b*) Bidimensionales: $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.
   - *c*) n-dimensionales: $(x_1^1, x_2^1, \cdots, x_n^1), \cdots, (x_1^r, x_2^r, \cdots, x_n^r)$.
4. **Por su dependencia temporal**
   - *a*) Temporales. Los valores se toman en instantes o períodos de tiempo.
   - *b*) Atemporales. No dependen de ningún soporte temporal.

2.4. Distribución de datos

La organización de los datos constituye la primera etapa de su tratamiento, pues, facilita los cálculos posteriores y evita posibles confusiones. Realmente, la organización de la información tiene una raiz histórica y aunque actualmente con el desarrollo de los medios informáticos deja de tener importancia desde un punto de vista aplicado, desde

la perspectiva de la enseñanza de la Estadística tiene un gran valor conceptual. La organización va a depender del número de observaciones distintas que se tengan y de las veces que se repitan cada una de ellas. En base a lo anterior se pueden estructurar los datos de tres maneras distintas:

1. *Tipo I:* Cuando se tiene un número pequeño de observaciones casi todas distintas, éstas se darán por extensión.

   Ejemplo 1.3 *En la serie: 2, 3, 5, 7, 7, 8, 11, 14, 16, 19, el 7 se repite dos veces y el resto de los valores está presente una vez.*

2. *Tipo II:* Cuando se tiene un gran número de observaciones pero muy pocas distintas, se organizan en una *tabla de frecuencias*, es decir, cada uno de los valores acompañado de la frecuencia con la que se presenta.

   Ejemplo 1.4 *La tabla*

Valor	*Frecuencia*
2	*4*
4	*4*
5	*3*
6	*2*
7	*3*
8	*3*
9	*1*

   indica que el valor 2 se repite 4 veces, el valor 4 se repite 4 veces, etc...

3. *Tipo III:* En el caso de que haya muchas observaciones, la mayoría de ellas distintas, pueden disponerse agrupándolas en intervalos e indicando el número de observaciones que caen dentro de cada intervalo.

Ejemplo 1.5 *La tabla*

Intervalo	*Frecuencia*
(2,3]	*4*
(3,7]	*6*
(7,12]	*12*
(12,21]	*8*
(21,25]	*6*
(25,30]	*4*
(30,50]	*3*

nos dice que en el intervalo $(2,3]$ *hay 4 observaciones, que en el* $(3,7]$ *hay 6, etc...*

En cualquiera de los tres casos o tipos se tiene una *distribución de frecuencias*. A la variable que representa a la distribución se le llama genéricamente X, a cada uno de los valores que toma la variable se le denota por x_i, y a la frecuencia con que toma dicho valor por n_i. Para evitar confusiones es aconsejable ordenar los valores de la variable de menor a mayor. Los valores ordenados de una distribución se presentan con los subíndices entre paréntesis:

$$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$$

de tal forma que siempre se verifica que $x_{(i)} \leq x_{(i+1)}$.

Para efectuar cálculos, sea cuál sea el tipo de distribución, se disponen los datos de la siguiente forma:

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_k	$N_r = n$	f_r	$F_r = 1$

Donde:

- n representa al número total de observaciones y será igual a $\sum_{i=1}^{r} n_i$

- f_i es la frecuencia relativa, definida como $\frac{n_i}{n}$
- N_i es la frecuencia absoluta acumulada, que se obtiene como $\sum_{j=1}^{i} n_j$
- F_i es la frecuencia relativa acumulada, que viene dada por $\sum_{j=1}^{i} f_j$

Observe que si la distribución es de tipo I cada una de las frecuencias absolutas es igual a 1, y si la distribución es de tipo III los valores x_i representan a las marcas de clase o puntos medios de los intervalos[1].

3. Representaciones gráficas

En función de la naturaleza de los datos y de la forma en que éstos se presenten existen distintos tipos de representaciones. Se muestran aquí las más interesantes.

1. El *diagrama de tarta* se emplea para representar atributos.

   Ejemplo 1.6 *En una votación entre cuatro candidatos a representante de una comunidad se han obtenido los siguientes resultados:*

Candidato	*Número de votos*
A	*287*
B	*315*
C	*275*
D	*189*

   La representación gráfica mediante un diagrama de tarta sería la que se muestra en la figura 1.1.

2. Una distribución dada por extensión, se representa mediante un *diagrama de puntos.*

[1] Dado el intervalo (L_i, L_{i+1}), la marca de clase viene dada por $x_i = \frac{L_i + L_{i+1}}{2}$

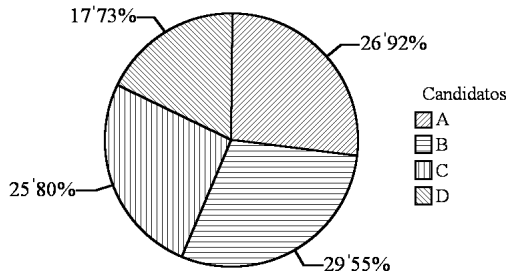

Figura 1.1: Diagrama de tarta

Ejemplo 1.7 *En un estudio sobre el peso y la estatura de un grupo de siete estudiantes se han obtenido las siguientes mediciones:*

$$(73, 1'87), (67, 1'75), (75, 1'80),$$
$$(66, 1'67), (80, 1'95), (64, 1'78), (83, 1'77).$$

La representación gráfica mediante un diagrama de puntos es la que se muestra en la figura 1.2.

A dicha representación se le suele denominar nube de puntos o diagrama de dispersión; se estudiará más a fondo en el capítulo 2.

3. Para representar una distribución del tipo II, se utiliza un *diagrama de barras*:

   Ejemplo 1.8 *La representación de la distribución del ejemplo 1.4 es la que se muestra en la figura 1.3.*

4. Por último, si se tiene una distribución del tipo III, se utiliza un *histograma*:

   Ejemplo 1.9 *El histograma correspondiente a la distribución del ejemplo 1.5 es el de la figura 1.4.*

   Observe que el efecto que produce el histograma es el de relacionar el número de observaciones con el área dentro de cada rectángulo, por lo que si éstos tienen la misma base, es decir, si los intervalos son de la misma amplitud, basta con construir rectángulos con base los intervalos y altura las frecuencias asociadas a ellos. En

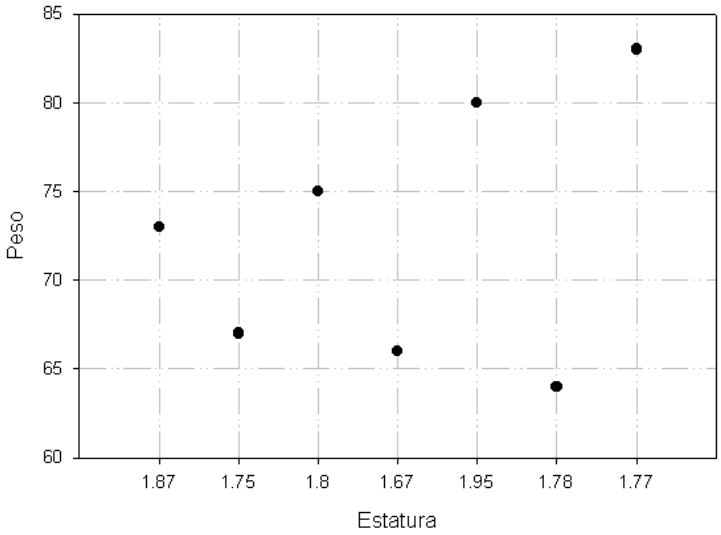

Figura 1.2: Diagrama de puntos

cambio, si las bases son distintas, o lo que es lo mismo, si los intervalos son de distinta amplitud, y se emplea el criterio anterior de asignación de alturas, se producirá una distorsión óptica. Por ello, en estos casos en vez de utilizar la frecuencia como altura de los rectángulos se utiliza la denominada *altura del histograma* o densidad de observaciones en el intervalo, definida como $h_i = \frac{n_i}{a_i}$, donde a_i es la amplitud del intervalo correspondiente.

4. Medidas centrales

Una vez organizados los datos en su correspondiente distribución de frecuencias, se procede a dar una serie de medidas que resuman toda esa información y que, "de alguna manera", representen a la distribución.

4.1. La media

La *media* es una medida de representación central que necesariamente debe cumplir tres requisitos:

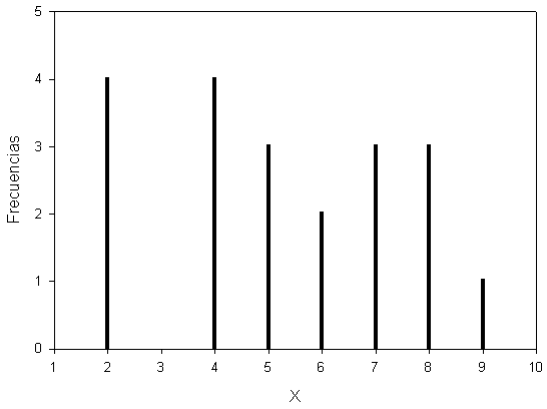

Figura 1.3: Diagrama de barras

1. Para su obtención deben utilizarse todas las observaciones.

2. Debe ser un valor comprendido entre el menor y el mayor de los valores de la distribución.

3. Debe venir expresada en la misma unidad que los datos.

Entre todas las funciones que verifican estas tres propiedades se destaca la *media aritmética*, a partir de ahora *media* simplemente, que se define de la siguiente manera:

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

Donde las x_i representan, según el caso, a los valores de la variable o a las marcas de clase de los intervalos.

Ejemplo 1.10 *La media de la distribución del ejemplo 1.4 viene dada por:*

$$\begin{aligned} \bar{x} &= \frac{2 \cdot 4 + 4 \cdot 4 + 5 \cdot 3 + \cdots + 9 \cdot 1}{4 + 4 + 3 + \cdots + 1} \\ &= \frac{105}{20} = 5'25. \end{aligned}$$

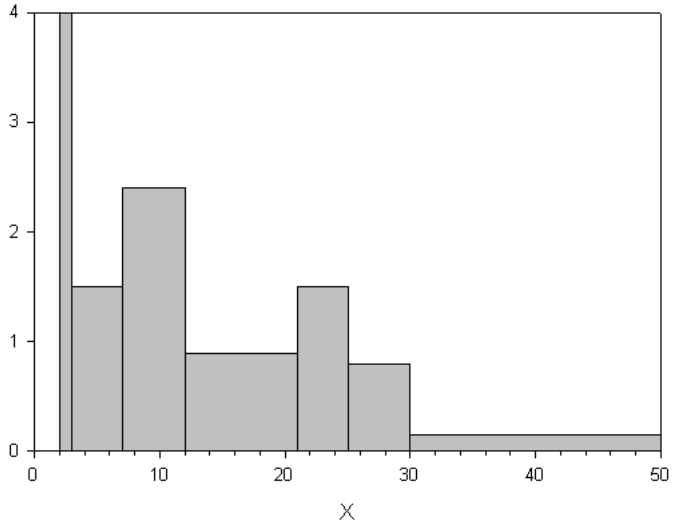

Figura 1.4: Histograma

Con el mismo esquema también se puede definir la *media geométrica* como:

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_r^{n_k}}.$$

Ejemplo 1.11 *La media geométrica de la distribución del ejemplo 1.3 se obtendría como:*

$$\bar{x}_g = \sqrt[10]{2 \cdot 3 \cdot 5 \cdot 7^2 \cdot \ldots \cdot 19} = 7'483.$$

Cuando se tiene que hacer un promedio de un grupo de razones se utiliza la *media armónica*, definida como:

$$\bar{x}_a = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}.$$

Ejemplo 1.12 *La media armónica de la distribución del ejemplo 1.4 se obtendría como:*

$$\bar{x}_a = \frac{20}{\frac{4}{2} + \frac{4}{4} + \frac{3}{5} + \cdots + \frac{1}{9}} = 4'125.$$

Otra media que tiene interés práctico es la *media ponderada*. Esta consiste en asignar a cada valor x_i un peso w_i que depende de la importancia relativa de cada uno de estos valores bajo algún criterio. Su expresión responde a:

$$\bar{x}_p = \frac{\sum_{i=1}^{r} n_i w_i x_i}{\sum_{i=1}^{r} n_i w_i}.$$

Ejemplo 1.13 *Para superar la asignatura de estadística, un alumno debe ser evaluado en distintas pruebas referentes a la misma: test, problemas y práctica, cada una de ellas ponderada según su importancia o contribución en la nota final. Así, los pesos de cada prueba serán del* 30 %, 50 % *y* 20 % *respectivamente. Sabiendo que las notas obtenidas por el alumno en cada prueba son 7, 3 y 5 respectivamente, ¿cuál es la nota global en la asignatura?*

$$\begin{aligned} \bar{x}_p &= \frac{7 \cdot 30 + 3 \cdot 50 + 5 \cdot 20}{30 + 50 + 20} \\ &= \frac{460}{100} = 4'6. \end{aligned}$$

Propiedades de la media. Se analizan a continuación una serie de propiedades de la media que hacen de ésta una medida óptima de representación.

1. La suma de las desviaciones de los valores de la distribución respecto a la media es igual a cero, es decir:

$$\sum_{i=1}^{r} (x_i - \bar{x}) n_i = 0.$$

2. Si a cada observación de una distribución X se le suma una constante k (traslación), se tiene una nueva variable $Y = X + k$ con media igual a la de X más la constante k.

3. Si se multiplica una variable X por una constante k (homotecia), la variable resultante $Y = kX$ tendrá media igual a k por la media de X.

   Estas dos propiedades se pueden resumir en la siguiente:

$$Y = aX + b \qquad \Rightarrow \qquad \bar{y} = a\bar{x} + b.$$

4. La media es el valor ϕ que hace mínima la expresión:

$$\sum_{i=1}^{r}(x_i - \phi)^2 n_i.$$

Precisamente ese mínimo será la varianza de X, medida de dispersión que se estudia más adelante. Por otra parte, se comprobará que esta propiedad de la media garantiza su bondad como medida de representación.

Ejercicio 1.4 *Demuestre las propiedades anteriores.*

4.2. La mediana

La *mediana* es un valor que, previa ordenación, deja la mitad de las observaciones en la recta real a la izquierda y la otra mitad a la derecha. Es decir, el 50 % de los datos son menores o iguales que la mediana y el otro 50 % mayores o iguales a ésta. Para su cálculo y suponiendo que los valores están ordenados se procede de la siguiente manera:

1. Si los datos vienen dados por extensión, y hay un número impar de ellos la mediana es el elemento que se encuentra en el centro, es decir $x_{(\frac{n+1}{2})}$. Si el número de datos fuera par habría dos elementos centrales y la mediana se obtendría como la media de ambos, es decir:

$$M_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Ejemplo 1.14 *La mediana de la distribución del ejemplo 1.3 se obtendría como:*

$$M_e = \frac{x_{(5)} + x_{(6)}}{2} = \frac{7+8}{2} = 7'5.$$

2. A partir de una distribución de tipo II ordenada, se construye la columna de frecuencias absolutas acumuladas, se obtiene el valor de $\frac{n}{2}$, deslizándose por la columna de N_i hasta detectar la primera frecuencia mayor o igual que $\frac{n}{2}$; si dicha frecuencia es estrictamente mayor que $\frac{n}{2}$ la mediana toma el valor de la observación que la ostenta, si por el contrario $\frac{n}{2}$ coincide con algún N_i la mediana vale $\frac{x_i+x_{i+1}}{2}$.

   Ejemplo 1.15 *Para calcular la mediana en la distribución del ejemplo 1.4 se obtiene $\frac{n}{2}$ que es igual a 10, construyendo la columna de frecuencias acumuladas:*

x_i	n_i	N_i	
2	4	4	
4	4	8	
5	3	11	$\longleftarrow$
6	2	13	
7	3	16	
8	3	19	
9	1	20	

   Puesto que $N_2 < 10$ y $N_3 > 10$ entonces $M_e = 5$.

3. Por último, si la distribución viene agrupada en intervalos, se construye también la columna de N_i para fijar el intervalo donde se halla la mediana, éste queda determinado porque es el primero que verifica que la frecuencia acumulada del intervalo es mayor o igual que $\frac{n}{2}$. Una vez fijado el intervalo, la mediana adopta la siguiente expresión:

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

   donde L_{i-1} es el extremo inferior del intervalo y a_i su amplitud.

Ejemplo 1.16 *En la distribución del ejemplo 1.5, $\frac{n}{2} = 21'5$. La tabla de frecuencias acumuladas que se obtiene es:*

$(L_{i-1}, L_i]$	n_i	N_i	
$(2, 3]$	4	4	
$(3, 7]$	6	10	
$(7, 12]$	12	22	$\longleftarrow$
$(12, 21]$	8	30	
$(21, 25]$	6	36	
$(25, 30]$	4	40	
$(30, 50]$	3	43	

Por tanto:

$$M_e = 7 + \frac{21'5 - 10}{12} 5 = 11'79.$$

Ejercicio 1.5 *Demuestre que la mediana es el valor ϕ que hace mínima la expresión:*

$$\sum_{i=1}^{r} |x_i - \phi| n_i.$$

4.3. Las modas

La *moda* absoluta de una distribución es el valor que más veces se repite. Además de la moda absoluta, aquellos valores que tengan frecuencia mayor a la de los valores adyacentes serán modas relativas.

Las modas se pueden obtener fácilmente cuando los datos vienen dados en forma puntual.

Ejemplo 1.17 *En la distribución 2, 3, 3, 4, 6, 7, 7, 7, 10, la moda absoluta es 7, puesto que es el valor que se repite más veces, concretamente 3. Además, el 3 es una moda relativa, puesto que su frecuencia es 2, superior a la de los valores 2 y 4, ambas iguales a 1.*

Si las observaciones vienen agrupadas en intervalos hay que distinguir dos casos:

1. Intervalos de igual amplitud. En este caso se fija el intervalo que tenga mayor frecuencia –intervalo modal absoluto– y aquellos con frecuencia superior a la de los intervalos adyacentes –intervalos modales relativos–. Dentro de cada intervalo modal la moda corresponde al valor:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i+1} + n_{i-1}} a_i.$$

Ejemplo 1.18 *En la distribución que sigue, el intervalo modal absoluto es el* $(4, 5]$*, además se tiene un intervalo modal relativo, el* $(6, 7]$.

$(L_{i-1}, L_i]$	n_i
$(2, 3]$	2
$(3, 4]$	3
$(4, 5]$	7
$(5, 6]$	3
$(6, 7]$	6
$(7, 8]$	5
$(8, 9]$	3

La moda absoluta será:

$$M_o = 4 + \frac{3}{3+3} 1 = 4'5.$$

Y la moda relativa:

$$M_o = 6 + \frac{5}{5+3} 1 = 6'625.$$

2. Intervalos de distinta amplitud. En este caso el intervalo modal absoluto será aquel que tenga mayor *altura de histograma*, h_i, con idéntica discusión que antes para las modas relativas. La expresión de la moda viene dada por:

$$M_o = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i.$$

Ejemplo 1.19 *Para la distribución que sigue:*

$(L_{i-1}, L_i]$	n_i	h_i
$(2, 3]$	1	1
$(3, 7]$	6	$1'5$
$(7, 9]$	12	6
$(9, 14]$	8	$1'6$
$(14, 20]$	6	1
$(20, 30]$	4	$0'4$

El intervalo modal, sólo existe uno, es $(7, 9]$*, con lo que la moda vale:*

$$M_o = 7 + \frac{1'6}{1'6 + 1'5} 2 = 8'032.$$

Para terminar este epígrafe observe que cuando las distribuciones son de intervalos los cálculos puntuales de la mediana y la moda utilizan criterios de ponderación que suponen, como no puede ser de otra manera, la disposición uniforme de las observaciones dentro de los intervalos.

4.4. Comparación entre media, moda y mediana

Salvo en casos muy específicos, la media es la mejor de las medidas de representación, pues la moda es bastante inestable y un pequeño cambio en las observaciones puede afectarle mucho, mientras que la mediana es insensible al tamaño de los datos, permaneciendo constante si, por ejemplo, se altera arbitrariamente y en cierto sentido las observaciones extremas. Por otra parte, si se dispone de las modas y medianas de dos distribuciones hay que conocer cada uno de los datos de éstas para calcular la moda y mediana de la distribución conjunta. La media por el contrario es sensible a las alteraciones de los datos, al tamaño de éstos y si se conocen las medias de dos conjuntos de datos, basta con saber los tamaños de ambos grupos para calcular la media global.

Ejercicio 1.6 *Calcule la media, mediana y moda de la distribución:*

1, 2, 4, 7, 9, 9, 9, 11, 13, 14, 17, 21, 34

Obtenga de nuevo dichas medidas para la distribución a la que se ha añadido los valores -1 y 47. Comente los resultados en lo que se refiere a la estabilidad de las medidas obtenidas.

5. Medidas de posición

Se llaman medidas de posición o *cuantiles* de orden k a aquellas que dividen a la distribución en k partes, de tal forma que en cada una de esas partes haya el mismo número de elementos[2]. De entre todas las medidas de posición destacan los *cuartiles*, los *deciles* y los *percentiles*. Los cuartiles dividen a la distribución en cuatro partes iguales, los deciles en diez y los percentiles en cien. Habrá, por tanto, tres cuartiles (Q_1, Q_2, Q_3), nueve deciles $(D_1, D_2, \cdots, D_9)$ y, noventa y nueve percentiles $(P_1, P_2, \cdots, P_{99})$. El segundo cuartil, el quinto decil y el quincuagésimo percentil son iguales y coinciden con la mediana. En distribuciones puntuales el cálculo es idéntico al de la mediana, siendo ahora $\frac{rn}{k}$ el valor de discusión. En distribuciones por intevalos la forma general de cálculo para un cuantil, al que se denota por $C_{\frac{rn}{k}}$, $k = 4, 10, 100, \ldots$, es la siguiente:

$$C_{\frac{rn}{k}} = L_{i-1} + \frac{\frac{rn}{k} - N_{i-1}}{n_i} a_i.$$

Siendo el intervalo i-ésimo el primero que verifica $N_i \geq \frac{rn}{k}$.

Ejemplo 1.20 *En la distribución:*

$(L_{i-1}, L_i]$	n_i	N_i	
$(2, 3]$	4	4	
$(3, 7]$	6	10	
$(7, 12]$	12	22	$\leftarrow P_{35}$
$(12, 21]$	8	30	
$(21, 25]$	6	36	$\leftarrow Q_3$
$(25, 30]$	4	40	
$(30, 50]$	3	43	

El Q_3 se obtendría calculando $\frac{3 \cdot 43}{4} = 32'5$. La primera frecuencia acumulada mayor que $32'5$ corres-

[2]La mediana es un caso particular de cuantil, que divide la distribución en dos partes iguales.

ponde al intervalo $(21, 25]$*, por lo que:*

$$Q_3 = 21 + \frac{32'5 - 30}{6}4 = 22'66.$$

Para calcular el P_{35} *se obtiene* $\frac{35 \cdot 43}{100} = 15'05$*. El intervalo donde se encuentra el percentil buscado es el* $(7, 12]$ *y, por tanto:*

$$P_{35} = 7 + \frac{15'05 - 10}{12}5 = 9'10.$$

6. Medidas de dispersión

A continuación se estudian una serie de medidas que por una parte indicarán el nivel de concentración de los datos que se están analizando y por otra informarán sobre la bondad de los promedios calculados como representantes del conjunto de datos.

6.1. Varianza y desviación típica

La *varianza* y su raíz cuadrada positiva, la *desviación típica*, son las más importantes medidas de dispersión, estando íntimamente ligadas a la media como medida de representación de ésta. La varianza viene dada por la expresión:

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

Y la desviación típica es, por tanto, $S = +\sqrt{S^2}$.

El dar dos expresiones para un mismo concepto se explica porque la varianza es un término de más fácil manejo, mientras que la desviación típica viene dada en la misma unidad que la variable. Tanto una como la otra son siempre positivas y valen cero sólo en el caso de que todos los valores coincidan con la media (representatividad absoluta de la media).

Ejemplo 1.21 *Dada la distribución:*

$(L_{i-1}, L_i]$	x_i	n_i
$(-2, 2]$	0	1
$(2, 4]$	3	3
$(4, 8]$	6	6
$(8, 12]$	10	13
$(12, 20]$	16	8
$(20, 24]$	22	6
$(24, 30]$	27	5
$(30, 40]$	35	3

Cuya media vale 15, se calcula la varianza y la desviación típica como:

$$S^2 = \frac{(0-15)^2 \cdot 1 + \cdots + (35-15)^2 \cdot 3}{45} = 82$$

$$S = \sqrt{82} = 9'055.$$

Propiedades de la varianza

1. Si se le suma una constante a una variable, la varianza de la nueva variable no cambia.

2. Si se multiplica una variable por una constante, la varianza de la nueva variable es igual a la de la antigua multiplicada por la constante al cuadrado.

Estas dos propiedades pueden resumirse en la siguiente expresión:

$$Y = aX + b \qquad \Rightarrow \qquad S_Y^2 = a^2 S_X^2.$$

Ejercicio 1.7 *Demuestre las propiedades anteriores.*

Ejemplo 1.22 *Dada la variable X con media $\bar{x} = 12$ y desviación típica $S_X = 9$, la variable $Y = 3X - 4$ tendrá de media y desviación típica:*

$$\bar{y} = 3\bar{x} - 4 = 3 \cdot 12 - 4 = 32$$

$$S_Y = \sqrt{3^2} \cdot S_X = \sqrt{9} \cdot 9 = 27.$$

6.2. Otras medidas de dispersión

6.2.1. El recorrido y el rango

Se define el primero como la diferencia entre el mayor y el menor de los valores y el segundo como el intervalo cuyos extremos son el mínimo y el máximo de la distribución. Tienen la ventaja de que son fáciles de calcular, aunque cuando hay valores aislados en las puntas de la distribución dan una visión distorsionada de la dispersión de ésta.

Ejemplo 1.23 *En la distribución del ejemplo 1.4 el recorrido vale 7, mientras que el rango es* $[2, 9]$.

6.2.2. La desviación absoluta

La desviación absoluta respecto a la media, está definida por:

$$D_m = \frac{\sum_{i=1}^{r} |x_i - \bar{x}| n_i}{n}.$$

También puede definirse respecto a la mediana, siendo ésta el valor que minimiza dicha expresión.

6.2.3. Recorrido intercuartílico

Viene dado por:

$$R_I = Q_3 - Q_1.$$

Es una medida adecuada para el caso en que se desee que determinadas observaciones extremas no intervengan, evitándose, de este modo, una

visión sesgada de la variabilidad de la distribución. Como inconveniente principal tiene que en su confección sólo intervienen el 50 % de los valores centrales.

Las expresiones que se acaban de ver expresan la dispersión de la distribución en términos absolutos, se precisa definir a partir de ellas, otras que hagan posible la comparación entre varias variables y que tengan en cuenta el tamaño de las observaciones. Obsérvese que la distribución formada por los elementos {0'1, 0'2, 0'3, 0'4, 0'5} y la que constituyen {1000'1, 1000'2, 1000'3, 1000'4, 1000'5} tienen la misma varianza y, sin embargo, es evidente que en el primero de los casos los elementos están muy dispersos y en el segundo bastante concentrados, ésto es consecuencia de la primera de las propiedades de la varianza. Para evitar estas situaciones se estudia la siguiente medida.

6.3. Coeficiente de variación

Se define como el cociente entre la desviación típica y el valor absoluto de la media. Se trata de una medida adimensional, tiene en cuenta el rango de valores en el que se mueve, permite comparar la dispersión de varias distribuciones, es invariante respecto a homotecias y sensible frente a traslaciones. Además de lo anterior, el *coeficiente de variación* da información sobre la representatividad de la media; y aunque no hay valores fijos de comparación, pues depende de circunstancias tales como el número de observaciones, se puede considerar, a efectos prácticos, una cota de $0'5$ como límite para admitir que la media representa aceptablemente al conjunto de la distribución.

Ejemplo 1.24 *En el caso del ejemplo 1.21 se tiene que:*

$$C_V = \frac{S}{|\bar{x}|} = \frac{9'055}{15} = 0'60.$$

Lo que implica que la media no representa en modo alguno al conjunto de la distribución.

6.4. Recorrido semiintercuartílico respecto a la mediana

Viene dado por:

$$R_{SI} = \frac{Q_3 - Q_1}{M_e}$$

que al igual que la anterior es una medida adimensional, con las ventajas e inconvenientes mencionados para el recorrido intercuartílico.

7. Desigualdad de Tchebychev

Esta desigualdad relaciona a la media y a la varianza y tiene la expresión:

$$f(|x_i - \bar{x}| \leq aS) \geq 1 - \frac{1}{a^2} \quad , \quad a > 1.$$

Que justifica el caracter de medida de dispersión de la varianza. Así, en un intervalo de centro la media y radio 4 veces la desviación típica se encuentra, al menos, el 93'75 por ciento de la distribución.

Observación 1.1 *La desigualdad de Tchebychev proporciona una cota inferior para el porcentaje de observaciones en un determinado intervalo con centro la media de la distribución.*

Ejemplo 1.25 *Dada una distribución con media, $\bar{x} = 25$, y desviación típica, $S = 4$, el intervalo $[\bar{x} - 3S, \bar{x} + 3S] = [13, 37]$ garantiza la presencia en su interior de, al menos, el $88'88\,\%$ de la distribución.*

8. Momentos de la distribución

Las medidas que se han visto hasta el momento presentan visiones parciales de la distribución, se pretende dar ahora una herramienta eficaz que generalice esa idea, de tal forma que la mayoría de las características se puedan expresar utilizando dicha herramienta. Así, se hace referencia a los *momentos de la distribución.*

8.1. Momentos respecto al origen

Se define el *momento de orden k respecto al origen* como:

$$a_k = \frac{\sum_{i=1}^{r} x_i^k n_i}{n}.$$

Es evidente que a_0 es igual a 1 y que a_1 es igual a la media.

8.2. Momentos respecto a la media

El *momento de orden k respecto a la media* viene dado por:

$$m_k = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^k n_i}{n}.$$

Se puede comprobar que m_0 es igual a 1, que m_1 es cero y que m_2 es la varianza.

Es posible expresar los momentos respecto a la media en función de los momentos respecto al origen.

Ejercicio 1.8 *Demuestre que:*

a) $m_2 = S^2 = a_2 - a_1^2$

y que:

b) $m_3 = a_3 - 3a_2a_1 + 2a_1^3$.

Ejemplo 1.26 *En el ejemplo 1.21 el cálculo de la varianza se podría haber hecho, utilizando la fórmula anterior, de la siguiente manera:*

$$\begin{aligned} S^2 &= \sum_{i=1}^{r} \frac{x_i^2 n_i}{n} - \bar{x}^2 \\ &= \frac{0^2 \cdot 1 + 3^2 \cdot 3 + \cdots + 35^2 \cdot 3}{45} - 15^2 \\ &= 82. \end{aligned}$$

9. Medidas de forma

Este epígrafe y el siguiente se detienen a analizar la "forma" de la distribución, tratando a la variable desde un enfoque distinto al seguido hasta ahora, en primer lugar se examina la simetría y a continuación el apuntamiento.

9.1. Simetría

Los *coeficientes de simetría* indicarán si la distribución es simétrica y, caso de no serlo, el tamaño y la tendencia de su asimetría. Para ello, se distinguen dos tipos de distribuciones, las que tienen forma de campana y las que no la tienen, empleándose expresiones alternativas para su cálculo.

1. Si la distribución tiene forma de campana se utiliza la expresión:

$$A_s = \frac{\bar{x} - M_o}{S}.$$

De tal forma que cuando A_s es igual a cero la distribución es simétrica, si es menor, asimétrica negativa o tendida a la derecha, y si es mayor, asimétrica positiva o tendida a la izquierda.

Ejemplo 1.27 *Dada la distribución campaniforme:*

$(L_i, L_{i+1}]$	x_i	n_i	h_i
$(2, 4]$	3	2	1
$(4, 8]$	6	6	$1'5$
$(8, 12]$	10	12	3
$(12, 20]$	16	12	$1'5$
$(20, 24]$	22	3	$0'75$

Donde $\bar{x} = 12$, $S = 5'12$ *y* $M_o = 10$, *ocurre que:*

$$A_s = \frac{12 - 10}{5'12} = 0'39.$$

La representación gráfica de la distribución viene dada en la figura 1.5.

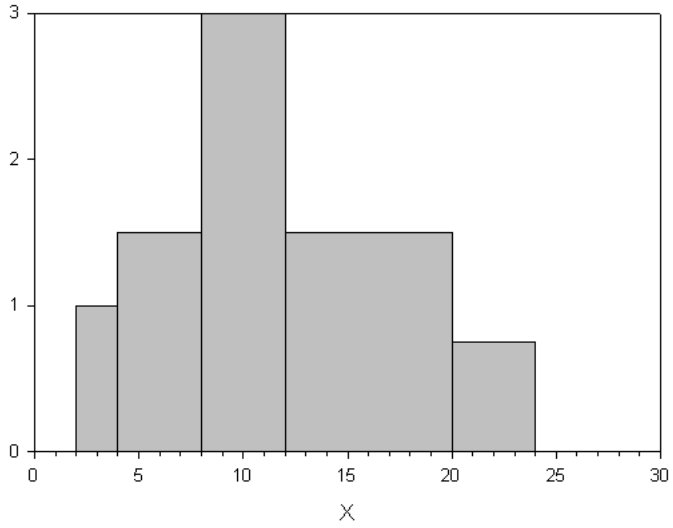

Figura 1.5: Histograma

Con lo que la distribución está, como puede observarse en el gráfico, inclinada, levemente, a la izquierda.

2. Si la distribución no tiene forma de campana o se desconoce este hecho se calcula la simetría mediante el coeficiente:

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

Siendo la discusión igual a la del caso anterior.

Observe que cuando la distribución es simétrica coinciden la media y la mediana, y que si además tiene forma de campana ambas son iguales a la moda.

9.2. Curtosis

El grado de apuntamiento de una distribución se examina a través del *coeficiente de curtosis*, para lo cual se compara con la distribución Normal tipificada o $N(0,1)$ que se trata en el capítulo 5 (figura 1.6).

Se puede adelantar, no obstante, que tiene forma de campana y que su estructura "probabilística" viene dada por la expresión:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

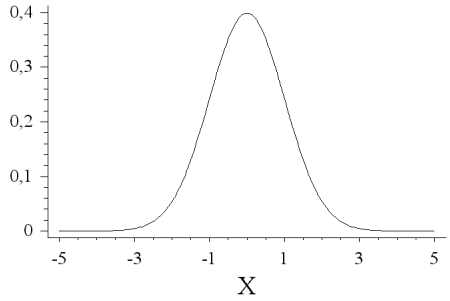

Figura 1.6: Función de densidad $N(0,1)$

El coeficiente de curtosis toma la expresión:

$$g_2 = \frac{m_4}{S^4}.$$

Cuando dicho coeficiente vale 3 coincide con el de la $N(0,1)$ y se dice que la distribución es mesocúrtica, si es menor que 3 platicúrtica y si es mayor que 3 leptocúrtica.

Ejemplo 1.28 *En la distribución de frecuencias:*

Valor	*Frecuencia*
2	*5*
4	*4*
5	*3*
6	*2*
7	*2*
8	*3*
9	*1*

claramente no campaniforme, se tiene que: $n = 20$, $\bar{x} = 5$, $S = 2'258$, $m_3 = 1'2$ *y* $m_4 = 47'1$ *por lo que*

el coeficiente de asimetría vendría dado por:

$$g_1 = \frac{m_3}{S^3} = \frac{1'2}{11'51} = 0'104.$$

Lo que implicaría que la distribución está levísimamente inclinada hacia la izquierda.

Por lo que respecta al coeficiente de curtosis:

$$g_2 = \frac{m_4}{S^4} = \frac{47'1}{25'99} = 1'81.$$

Tratándose, por consiguiente, de una distribución claramente aplastada o platicúrtica.

10. Transformaciones

A veces se tiene el inconveniente de que la distribución que se estudia presenta muchas irregularidades, como asimetrías acentuadas, valores extremos, etc..., en otras ocasiones se debe comparar la posición de dos elementos que pertenecen a poblaciones con características distintas o del mismo elemento en situaciones distintas. En estos casos es recomendable efectuar una transformación que haga más regular la distribución y, por tanto, con mejores condiciones para su estudio. Particular importancia tiene la tipificación de una variable.

10.1. Normalización o tipificación

Dada una variable X con media $\bar{x}$ y desviación típica S, la tipificación consiste en realizar la siguiente transformación:

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

A la nueva variable Z se le llama *variable normalizada* o *tipificada* y tiene media 0 y desviación típica 1. Haciendo un símil, la media y la desviación típica de una variable pueden considerarse como el centro de gravedad de la distribución y su escala, respectivamente, por lo que al tipificar distintas variables las centramos en el mismo punto y las

dotamos de la misma escala; además, los valores tipificados pierden la unidad de la variable. Por lo anterior, la tipificación tiene la propiedad de hacer comparables individuos que pertenezcan a distintas distribuciones, aún en el caso de que éstas vinieran expresadas en diferentes unidades.

Ejemplo 1.29 *Dos trabajadores del mismo sector ganan 620€ y 672€, respectivamente. El primero pertenece a la empresa A, cuya retribución media y desviación típica vienen dados por:* $\bar{x}_A = 580$€ *y* $S_{x_A} = 25$€*, mientras que para la empresa del segundo trabajador se tiene:* $\bar{x}_B = 640$€ *y* $S_{x_B} = 33$€*. Tanto uno como el otro ganan salarios por encima de la media, por lo que si se quiere conocer cuál de los dos ocupa mejor posición relativa dentro de su empresa hay que tipificar sus puntuaciones, y así:*

$$z_A = \frac{620 - 580}{25} = 1'6$$

mientras que:

$$z_B = \frac{672 - 640}{33} = 0'97.$$

Por lo que, aunque en términos absolutos el trabajador de la empresa B gana más que el de A, en relación al conjunto de los empleados de cada empresa el empleado de A ocupa mejor posición.

Otras transformaciones usuales son la del logaritmo y la de la raíz cuadrada que consiguen una mayor simetría y concentración de los valores de la distribución.

11. Análisis exploratorio de datos

El análisis exploratorio de datos (AED) está formado por un conjunto de técnicas estadísticas, fundamentalmente gráficas, que pretenden dar una visión simple e intuitiva de las principales características de la distribución en estudio. El AED puede ser un fin por sí mismo o una primera etapa de un estudio más completo. Como aspectos más desta-

cables que abarca el AED, están los que se refieren a la forma de la distribución y a la detección de valores anómalos.

11.1. Diagramas de tallo y hojas de Tukey

El *diagrama de tallo y hojas* es una representación semi-gráfica donde se muestra el rango y distribución de los datos, la simetría y si hay candidatos a valores atípicos. Para su construcción se siguen los siguientes pasos:

1. Se redondean los valores a dos o tres cifras significativas.
2. Se divide el rango de los datos en k intervalos, cada uno representado por una fila de la tabla que está dividida por una línea vertical en dos partes. En cada fila, los datos individuales son representados por uno o dos dígitos, según el rango, (llamado tallo), mientras que a la derecha de la línea vertical se coloca el último dígito del valor (llamado hoja). Si hay algún punto que se encuentra lejano de la mayoría de los valores (candidato a valor atípico), éste es colocado en hoja superior o inferior separada. La tabla de tallo y hojas se acompaña de una columna de frecuencias acumuladas creciente inferior y superiormente hasta el tallo que contiene la mediana que queda señalado entre paréntesis.

Ejemplo 1.30 *A partir de la información recogida sobre los caballos de potencia de distintos vehículos, se representa el diagrama de tallo y hojas para dicha variable (figura 1.7).*

Su uso es recomendable siempre que el número de datos no sea muy grande (menor que 50).

11.2. Diagrama de caja ó diagrama de box-whisker

Los *diagramas de caja* son representaciones gráficas sencillas que no necesitan un número elevado de valores para su construcción. Se utilizan para estudiar tanto la dispersión como la forma de una distribución.

```
unidad = 10,0  1|2  representa 120,0

  1  0|5
  7  0|666777
 26  0|8889999999999999999
(20) 1|00000000001111111111
 43  1|2233333
 36  1|44444444444555555
 19  1|6666666677
  9  1|889
  6  2|000
  3  2|22

 HI | 245,0
```

Figura 1.7: Diagrama de tallo y hojas

Asimismo son especialmente útiles para comparar distintas distribuciones entre sí.

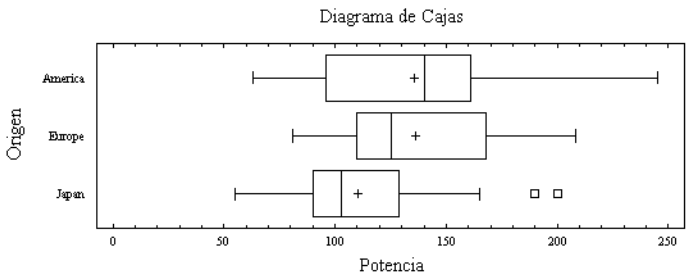

Figura 1.8: Diagrama de cajas

La caja representa el 50 % central de la distribución, la línea situada en el interior de la caja es la mediana, mientras que la cruz se corresponde con la media. Los extremos inferiores y superiores de los segmentos (también llamados bigotes) delimitan lo que se denomina como valores "normales" y coinciden, respectivamente, con el mínimo y el máximo de los valores una vez excluidos los candidatos a valores anómalos. Los candidatos a valores anómalos se etiquetan como atípicos y coinciden con aquellas observaciones que se encuentran fuera del

intervalo (LI, LS), donde:

$$\begin{aligned} LI &= Q_1 - 1'5R_I \\ LS &= Q_3 + 1'5R_I, \end{aligned}$$

es decir, a una distancia de Q_1, por la izquierda, o de Q_3, por la derecha, superior a una vez y media el recorrido intercuartílico; denominándose, en este caso, atípicos de primer nivel. Cuando la distancia, por uno de los dos lados, es superior a tres recorridos intercuartílicos, el valor atípico se denomina de segundo nivel.

Los valores atípicos de primer y segundo nivel quedan normalmente identificados en el diagrama de cajas por símbolos diferenciados ($\triangle$, $\diamondsuit$, $\cdot$), debiendo considerarse la posibilidad de realizar una depuración de los mismos antes de comenzar el tratamiento de los datos.

12. Ejercicios

12.1. Ejercicio resuelto

1.1 Para realizar un determinado experimento se ha medido la anchura interorbital, en mm., de una muestra de 40 palomas, obteniéndose los siguientes datos:

12'2, 12'9, 11'8, 11'9, 11'6, 11'1, 12'3, 12'2, 11'8, 11'8
10'7, 11'5, 11'3, 11'2, 11'6, 11'9, 13'3, 11'2, 10'5, 11'1
12'1, 11'9, 10'4, 10'7, 10'8, 11'0, 11'9, 10'2, 10'9, 11'6
10'8, 11'6, 10'4, 10'7, 12'0, 12'4, 11'7, 11'8, 11'3, 11'1

Se pide:

a) Construya una distribución de frecuencias y calcule la media, desviación típica y coeficiente de variación.

b) Agrupe los datos en intervalos con la amplitud más adecuada, calculando de nuevo los parámetros anteriores y comparándolos con los resultados obtenidos a partir de los datos no agrupados. Dibuje el histograma.

En lo que sigue trabaje con la distribución por intervalos.

c) ¿En qué intervalo de centro la media se encuentra, al menos, el 75 % de la distribución?

d) Calcule la mediana y la moda.

e) Obtenga el intervalo donde se encuentra el 40 % central de la distribución.

f) Estudie la simetría y el apuntamiento de la distribución.

Solución:

a) La distribución de frecuencias sería:

x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
$10'2$	1	$11'0$	1	$11'7$	1	$12'2$	2
$10'4$	2	$11'1$	3	$11'8$	4	$12'3$	1
$10'5$	1	$11'2$	2	$11'9$	4	$12'4$	1
$10'7$	3	$11'3$	2	$12'0$	1	$12'9$	1
$10'8$	2	$11'5$	1	$12'1$	1	$13'3$	1
$10'9$	1	$11'6$	4				

Gráficamente dicha distribución puede presentarse mediante el polígono de frecuencias de la figura 1.9.

Para calcular la media:

$$\bar{x} = \sum_{i=1}^{r} \frac{x_i n_i}{n} = \frac{459'2}{40} = 11'48 \text{ mm}$$

Es conveniente comprobar siempre que la media es un valor razonable y, en particular, dentro del rango de valores de la variable. En nuestro caso $10'2 < 11'48 < 13'3$.

La desviación típica vendría dada por:

$$\begin{aligned} S &= \sqrt{\sum_{i=1}^{r} \frac{x_i^2 n_i}{n} - \bar{x}^2} \\ &= \sqrt{\frac{5290'28}{40} - (11'48)^2} \\ &= \sqrt{0'4666} = 0'6831 \text{ mm} \end{aligned}$$

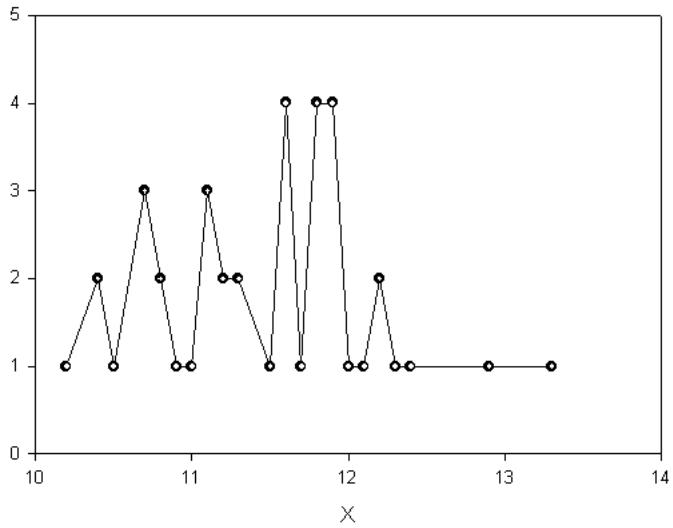

Figura 1.9: Polígono de frecuencias

Y el coeficiente de variación:

$$CV = \frac{S}{|\bar{x}|} = \frac{0'6831}{11'48} = 0'0595.$$

El bajo valor del coeficiente de variación indica que los valores están muy concentrados y que la media representa aceptablemente al conjunto de la distribución. En general, valores de CV menores a $0'1$ indican una alta concentración, entre $0'1$ y $0'5$ una concentración media y valores superiores a $0'5$ una alta dispersión y una media poco o nada representativa.

Observe que tanto la desviación típica como el coeficiente de variación son medidas positivas.

b) Para agrupar la distribución en intervalos se elige un número de éstos alrededor de $\sqrt{n}$, en nuestro caso $\sqrt{40} = 6'32 \simeq 7$. Los intervalos son de amplitud aproximada:

$$\frac{\text{Recorrido}}{\text{Nº de intervalos}} = \frac{13'3 - 10'2}{7} = 0'44.$$

Buscando siempre que sea un valor fácil de manejar, en este caso se opta por una amplitud de $0'5$. La distribución en intervalos quedaría:

x_i	$[L_{i-1}, L_i)$	n_i
$10'25$	$[10, 10'5)$	3
$10'75$	$[10'5, 11)$	7
$11'25$	$[11, 11'5)$	8
$11'75$	$[11'5, 12)$	14
$12'25$	$[12, 12'5)$	6
$12'75$	$[12'5, 13)$	1
$13'25$	$[13, 13'5)$	1

donde ahora x_i representa la marca de clase.

A partir de estos datos se tiene: $\bar{x} = 11'5$mm, $S = 0'6708$mm y $CV = 0'0583$. Con pequeñas variaciones respecto a los valores obtenidos para la distribución original, en todo caso, perfectamente asimilables y habiéndose conseguido una mayor facilidad de cálculo.

El histograma se representa en la figura 1.10. Como se puede apreciar la información visual que proporciona es mucho más clara que la que daría el polígono de frecuencias, a todas luces ininteligible. Se trata de una distribución unimodal y un poco tendida hacia la derecha, aunque esto se cuantificará más adelante.

c) Para contestar a esta cuestión se utiliza la desigualdad de Tchebychev, que dice:

$$f(|x_i - \bar{x}| \leq kS) \geq 1 - \frac{1}{k^2}.$$

Para $k = 2$, $1 - \frac{1}{k^2} = 0'75$, por lo que operando con el valor absoluto, el intervalo será:

$$[\bar{x} - 2S, \bar{x} + 2S] = [10'1138, 12'8462].$$

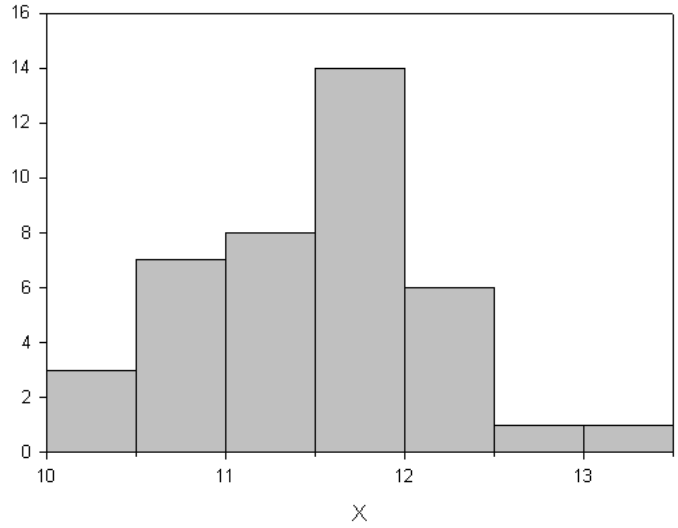

Figura 1.10: Histograma

d) Para calcular la mediana se obtiene la columna de frecuencias acumuladas:

x_i	$(L_{i-1}, L_i]$	n_i	N_i	
$10'25$	$10 - 10'5$	3	3	
$10'75$	$10'5 - 11$	7	10	
$11'25$	$11 - 11'5$	8	18	
$11'75$	$11'5 - 12$	14	32	$\leftarrow$
$12'25$	$12 - 12'5$	6	38	
$12'75$	$12'5 - 13$	1	39	
$13'25$	$13 - 13'5$	1	40	

La mediana se encuentra en aquel intervalo tal que $N_i \geq \frac{n}{2} = 20$, por tanto $M_e \in (11'5, 12]$, por lo que utilizando la fórmula apropiada, se tiene:

$$M_e = L_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i = 11'5 + \frac{20 - 18}{14} 0'5 = 11'5714 \text{ mm}$$

Por lo que $11'5714$ deja el $50\,\%$ de la distribución a la izquierda y el otro $50\,\%$ a la derecha.

Para calcular la moda, puesto que los intervalos son de igual amplitud, se selecciona aquel que tenga mayor frecuencia, en este caso el $(11'5, 12]$ que tiene frecuencia 14, y se aplica la fórmula correspondiente:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i = 11'5 + \frac{6}{6+8} 0'5 = 11'7143 \text{ mm}$$

e) El 40 % central de la distribución está contenido en el intervalo (P_{30}, P_{70}). El percentil P_{30} se encuentra en el intervalo $(L_{i-1}, L_i]$ para el que se verifica que $N_i \geq \frac{30 \cdot 40}{100} = 12$. Observando la columna de frecuencias acumuladas se ve que dicho intervalo es el $(11, 11'5]$. Por tanto:

$$P_{30} = 11 + \frac{12 - 10}{8} 0'5 = 11'125.$$

Operando de forma análoga:

$$P_{70} = 11'5 + \frac{28 - 12}{14} 0'5 = 11'8571.$$

Por lo que el intervalo pedido será el $(11'125, 11'8571)$.

f) Puesto que la distribución tiene forma de campana el coeficiente de simetría viene dado por:

$$A_s = \frac{\bar{x} - M_o}{S} = \frac{11'5 - 11'7143}{0'6708} = -0'319.$$

Por lo que la distribución está ligeramente inclinada hacia la derecha.

El coeficiente de apuntamiento:

$$g_2 = \frac{m_4}{S^4} = \frac{\frac{1}{40} \sum_{i=1}^{7} (x_i - \bar{x})^4 f_i}{(0'6708)^4} = \frac{0'400488}{0'202475} = 1'97796.$$

Al ser $g_2 < 3$ la distribución es platicúrtica, es decir, más aplastada que la distribución $N(0, 1)$.

12.2. Ejercicios propuestos

1.1. Al comenzar el curso se pasó una encuesta a los alumnos del primer curso de un colegio, preguntándoles, entre otras cuestiones, por el número de hermanos que tenían, obteniéndose los siguientes resultados:

3, 3, 2, 2, 8, 5, 2, 4, 3, 1, 4, 5, 3, 3, 3, 3, 3, 2, 5
1, 3, 3, 2, 2, 4, 3, 3, 2, 2, 4, 4, 3, 6, 3, 3, 2, 2, 4
3, 4, 3, 2, 2, 4, 4, 3, 3, 4, 2, 5, 4, 1, 2, 8, 2 ,3, 3, 4

a) Represente este conjunto de datos con un diagrama de barras.

b) Calcule media, moda y mediana.

c) Estudie la dispersión de los datos.

d) Analice la simetría de la distribución.

1.2. Los pesos de un colectivo de niños son:

60, 56, 54, 48, 99, 65, 58, 55, 74, 52, 53, 58, 67, 62, 65
76, 85, 92, 66, 62, 73, 66, 59, 57, 54, 53, 58, 57, 55, 60
65, 65, 74, 55, 73, 97, 82, 80, 64, 70, 101, 72, 96, 73, 55
59, 67, 49, 90, 58, 63, 96, 100, 70, 53, 67, 60, 54

Obtenga:

a) La distribución de frecuencias agrupando por intervalos.

b) La mediana de la distribución.

c) La media de la distribución, indicando su nivel de representatividad.

d) Utilizando la agrupación en intervalos, el porcentaje de alumnos que tienen un peso menor de 65 kg y el número de alumnos con un peso mayor de 60 kg dentro del grupo de los que pesan menos de 80 kg.

1.3. En el Consejo de Apuestas del Estado se han ido anotando, durante una temporada, el número de premiados de quinielas según la cantidad de aciertos, obteniéndose la siguiente tabla:

Nº de aciertos	11	12	13	14	15
Nº de personas (miles)	52	820	572	215	41

Calcule:

a) La mediana, la moda y los cuartiles de la distribución.

b) La simetría de la distribución.

1.4. En un puerto se controla diariamente la entrada de pesqueros según su tonelaje, resultando para un cierto día los siguientes datos:

Peso(Tm.)	0-25	25-50	50-70	70-100	100-500
Nº de barcos	5	17	30	25	3

Se pide:

a) El peso medio de los barcos que entran en el puerto diariamente, indicando la representatividad de dicha medida.

b) El intervalo donde se encuentra el 60 % central de la distribución.

c) El grado de apuntamiento.

d) El tonelaje más frecuente en este puerto.

1.5. El número de días de hospitalización de los enfermos que llegan en un cierto día a un servicio de urgencias, viene dado por:

Nº de días	0-1	2-5	6-8	9-15
Nº de enfermos	53	24	16	7

Se pide:

a) Un coeficiente que represente la distribución indicando dicho nivel de representatividad.

b) El porcentaje de enfermos que se quedan hospitalizados más de 5 días.

c) El valor que divide a la distribución en dos partes iguales.

1.6. Según un estudio se sabe que la planificación óptima de una determinada empresa exige que el 70 % sean administrativos, el 25 % jefes de departamento y el 5 % inspectores. Para realizar esta planificación se lleva a cabo un examen tipo test, obteniéndose las siguientes puntua-

ciones:

Puntuación	Empleados
[0,20)	70
[20,50)	115
[50,75)	95
[75,100)	5

a) ¿Cúal es la puntuación mínima para ser jefe de departamento?

b) ¿Y para ser inspector?

1.7. Para la selección de personal en dos empresas se realiza un test obteniéndose las siguientes puntuaciones porcentuales:

Factoría I		Factoría II	
Puntuación	Porcentaje	Puntuación	Porcentaje
[0,10]	0'07	[10,20]	0'08
[11,19]	0'25	[21,25]	0'16
[20,28]	0'38	[26,30]	0'20
[29,41]	0'19	[31,39]	0'28
[42,50]	0'11	[40,44]	0'23
		[45,50]	0'05

a) ¿Cuál de las dos factorías ha sido menos homogénea en los resultados?

b) ¿Qué persona ha tenido una puntuación mayor con respecto a su factoría: el que ha obtenido 35 en la factoría I o el que consiguió 38 en la II?

1.8. La vida útil de cierto tipo de bombonas de gas presenta la siguiente distribución:

Horas	Fracción de bombonas
[10,30)	0'04
[30,40)	0'27
[40,50)	0'34
[50,70)	0'26
[70,80]	0'09

Calcule:

a) La vida media de las bombonas de gas.

b) El tiempo de vida más frecuente.

c) El intervalo, con centro la media, donde se encuentre, al menos, el 85 % de la distribución.

d) El apuntamiento de la distribución.

1.9. El gasto de 100 experimentos, siendo la unidad 100€, viene dado por la siguiente tabla:

Gasto	[10,20)	[20,30)	[30,40)	[40,55]
Nº Experimentos	15	50	30	5

Calcule:

a) El gasto medio de los experimentos.

b) El porcentaje de experimentos que tienen un gasto entre 2300€ y 3500€.

c) Los precios que dividen a la distribución en cuatro partes iguales.

d) El gasto más frecuente.

1.10. En una entidad bancaria se sabe que, por término medio, el 15 % de los cheques son sin fondo. Las cantidades recogidas en dichos cheques, en euros, son las siguientes:

Importe de los cheques	Número de cheques
[0,200)	325
[200,600)	515
[600,1000)	420
[1000,3000]	270

Calcule:

a) El importe medio de los cheques sin fondo.

b) El importe más frecuente de los cheques sin fondo.

1.11. Para un determinado experimento se venía trabajando con unas temperaturas que variaban entre 100ºC y 130ºC. Estas temperaturas tenían una media de 110ºC y una desviación típica de 16ºC. Con un nuevo sistema se ha conseguido aumentar esta temperatura en 12ºC. ¿Cómo varía la dispersión relativa de dicha temperatura?

1.12. La producción de una empresa está organizada en dos factorías. La distribución de los salarios en cada una de ellas es la siguiente:

Salario en euros	Obreros Factoría A	Obreros Factoría B
[180,360)	20	20
[360,480)	23	28
[480,600)	22	14
[600,900)	15	8
[900,1200]	3	2

Se pide:

a) El salario más frecuente de la factoría A.

b) El salario que divide a la distribución de la factoría B en dos trozos iguales.

c) Los salarios medios de cada factoría.

d) El salario medio total a partir de los salarios medios de cada factoría.

1.13. El consumo de gasolina de dos coches de las marcas Citroën y Mercedes es, respectivamente, de 10 y 11 litros cada 100 km. Para el conjunto de los coches Citroën y Mercedes, se tienen, respectivamente, consumos medios de $7'4$ y $10'5$ litros y varianzas de 9 y 16 litros^2. Indique cuál de los dos coches tiene mayor consumo relativo dentro de su grupo.

1.14. En los contratos de venta de un fabricante, existe una cláusula por la que acepta la devolución de piezas defectuosas. Se consideran defectuosas aquellas cuya longitud no esté comprendida entre $(\bar{x}-l, \bar{x}+l)$. La longitud media de dichas piezas es de 22 mm. con desviación típica $0'3$. ¿Cuánto debe valer l para que el porcentaje de piezas devueltas no supere el 10 %?

1.15. Una empresa automovilística ha realizado un estudio sobre el grado de satisfacción de sus clientes (X) con la compra de vehículos pertenecientes a los segmentos medio (M) y alto (A), obteniendo los siguientes resultados:

X	n_M	n_A
0-6	4	4
7-13	6	7
14-20	9	9
21-27	12	8
28-34	9	2

a) Se considera que el grado de satisfacción es aceptable si la puntuación obtenida es superior a 19. Calcule el porcentaje de personas en cada grupo con un grado de satisfacción aceptable.

b) ¿Cuál de los dos grupos presenta mayor variabilidad?

1.16. Una población está dividida en dos subpoblaciones A y B, de las cuales se conoce lo siguiente:

$$n_A = 12,\ n_B = 9,\ \sum_A x_i = 234, \sum_B x_i = 138,\ \sum_A x_i^2 = 5036,$$

$$\sum_B x_i^2 = 2586,\ M_{O_A} = 20,\ M_{O_B} = 22,\ M_{e_A} = 19,\ M_{e_B} = 16.$$

a) ¿Se puede calcular la media global del conjunto? En caso afirmativo, calcúlela.

b) ¿Se puede calcular la moda y la mediana global del conjunto? En caso afirmativo, calcúlela.

c) ¿Cuál de las medias de las dos subpoblaciones es más representativa?

1.17. De una distribución se sabe que su media vale 5 y que el momento de orden dos con respecto al origen vale 29. Obtenga una cota inferior del porcentaje de dicha distribución que se encuentra en el intervalo $[2, 8]$.

1.18. Cuantificando a seis individuos en las variables X e Y, se dispone sólo de algunos de estos valores, se muestra a continuación la información disponible:

X	16	6	1		6	6
Y				24		

Complete la tabla sabiendo que la media de Y vale 14, que $\sum_{i=1}^{6} x_i = 48$, y que al tipificar los valores de X se obtiene el mismo resultado que al tipificar los valores de Y.