ANÁLISIS DESCRIPTIVO DE VARIABLES CUANTITATIVAS (2)

Contenidos

4.4. Introducción
4.5. Distribuciones de frecuencias de variables cuantitativas (datos agrupados)
4.6. Propiedades de las distribuciones de variables cuantitativas en muestras grandes
4.7. Variables discretas
4.7.1.Herramientas para el análisis gráfico
4.7.1.1. Tablas de distribuciones de frecuencia
4.7.1.2. Representaciones gráficas
4.7.2.Herramientas para el análisis numérico (estadígrafos)
4.7.2.1. Medidas de posición
4.7.2.2. Medidas de dispersión
4.7.2.3. Medidas de forma: asimetría y curtosis
4.8. Variables continuas
4.8.1. Herramientas para el análisis gráfico
4.8.1.1. Tablas de distribuciones de frecuencias
4.8.1.2. Representación gráfica
4.8.2. Herramientas para el análisis numérico (estadígrafos)
4.8.2.1.Medidas de posición
4.8.2.2.Medidas de dispersión
4.8.2.3. Medidas de forma: asimetría y curtosis
4.9. Comunicación y presentación de resultados

4.4. INTRODUCCIÓN

Con este capítulo se cierra la presentación de la Unidad I, destinada al aprendizaje del análisis estadístico descriptivo. Se espera que al alumno le haya quedado en claro que este tipo de análisis se inicia con una primera fase que persigue la organización y depuración de los datos, continúa con una segunda donde se aplican herramientas de análisis gráfico y numérico, y concluye con la elaboración de una comunicación que contiene un resumen sobre la metodología, y la interpretación de los resultados acompañada de elementos de presentación (tablas y representaciones gráficas significativas). Además debería quedar clara la etapa preparatoria del análisis de datos, donde se deberá ser cuidadoso para elegir las herramientas que corresponden a las diferentes situaciones (tipo de variable y tamaño muestral).

Cuando se dispone de una muestra pequeña de datos cuantitativos se ha visto que el análisis gráfico y numérico se aplica sobre una distribución simple de frecuencias. Particularmente en este capítulo se presentará el análisis gráfico y numérico relacionado con distribuciones de frecuencias de datos agrupados, referidos a los dos tipos de variables cuantitativas: discreta y continua. Dado que al tratarse de variables que en muestras de tamaño grande pueden tomar numerosos valores de la variable, la etapa inicial del análisis descriptivo estará destinada a obtener tales distribuciones de frecuencias, y la siguiente etapa a aplicar las herramientas gráficas y numéricas que en este caso presentan muchas posibilidades.

4.5. DISTRIBUCIONES DE FRECUENCIAS DE VARIABLES CUANTITATIVAS

En muestras grandes, el objetivo de la organización, esencialmente es resumir la cantidad de datos. El criterio a aplicar es: a) agrupar los datos en clases cualitativas o numéricas y, b) contar la cantidad de datos que resulta clasificado en cada grupo; esos conteos reciben el nombre de frecuencias. La serie completa de clases puestas en correspondencia con los conteos o frecuencias, se denomina **distribución de frecuencias**.

El término frecuencias es de carácter general, según el objetivo, será el tipo de frecuencias que utilicemos: **frecuencias absolutas, frecuencias relativas, frecuencias acumuladas o frecuencias expresadas en porcentaje**.

Las distribuciones de frecuencias de variables cualitativas y cuantitativas pueden ser presentadas en forma analítica a través de una **tabla de distribución de frecuencias**, o bien en forma gráfica a través de representaciones gráficas. En este último caso los gráficos son diferenciados. Cuando la variable es cualitativa se utilizarán: **diagramas de sectores y diagramas de barras**. A las variables

cuantitativas se les aplicará: a) **diagramas de frecuencias** o diagramas de líneas (variables discretas) o b) gráficos varios: **histograma, polígono de frecuencias o polígonos de frecuencias acumuladas** (variables continuas).

Con las distribuciones de frecuencias, puede decirse, que se cumple la primer etapa del proceso de dar sentido a los datos. Una distribución de frecuencias pone en evidencia a diversos aspectos sumamente importantes, referidos a las propiedades de los datos en masa, que permiten comprender el comportamiento de las variables, las cuales en el capítulo siguiente serán cuantificadas mediante las correspondientes medidas descriptivas o estadígrafos.

Resulta conveniente recordar la estructura que poseen las tablas utilizadas para sintetizar la clasificación de una muestra de tamaño n, en el caso de tener los datos de una variable cualitativa y de una cuantitativa (discreta y continua), a través del Cuadro 4.1.

Cuadro 4.1. Síntesis comparativa de la estructura de los datos agrupados según tipo de variable

Caso: Distribución de una variable cualitativa (clases categóricas)		Caso: Distribución de una variable cuantitativa (clases numéricas)			
		Tipo I		Tipo II	
Clase (a_i)	Conteo (n_i)	Valor observado de la variable, (x_i)	Conteo (n_i)	Intervalos de Clases	Conteo (n_i)
a_1	n_1	x_1	n_1	$[x_1 , x_2)$	n_1
a_2	n_2	x_2	n_2	$[x_2 , x_3)$	n_2
.	.	.	.	.	.
.	.	.	.	.	.
a_k	n_k	x_k	n_k	$[x_{k-1} , x_k)$	n_k

En todos los casos el conteo hace referencia al número de observaciones o mediciones clasificadas en la clase i-ésima de una variable. En el caso de variables cuantitativas discretas esa clase es de tipo puntual (valor puntual) mientras que en variables continuas se trata de un intervalo de valores. Con la claridad de este significado, se pasará a formalizar algunos conceptos frecuentistas.

Definición 4.15.
La serie de clases (cualitativas o cuantitativas) asociadas a sus correspondientes frecuencias, se llama **distribución de frecuencias**, e indica como la frecuencia total o cantidad total de datos se reparte entre los k agrupamientos realizados.
Según el tipo de frecuencia considerada se tendrá una distribución de frecuencias (absolutas), una distribución de frecuencias relativas o una distribución de frecuencias acumuladas. Cualquiera de ellas, se puede presentar tanto en forma tabular como gráficamente.

Definición 4.16
En datos agrupados, la **frecuencia absoluta** de una clase (cualitativa o cuantitativa), o simplemente frecuencia, simbolizada con n_i, está dada por el número de unidades de análisis clasificado en la clase i-ésima. La serie de frecuencias absolutas, para las k clases, se indica como
$n_1, n_2,\ldots,n_k$ tanto en el caso de datos categóricos como cuantitativos
Es fácil notar que las frecuencias absolutas cumplen con la siguiente propiedad: $n = n_1+n_2+\ldots+n_k$, por tanto
$n = \sum_{i=1}^{k} n_i$,
es decir, la suma total de las frecuencias absolutas es igual al tamaño muestral.

Definición 4.17.
La proporción dada por el cociente entre la frecuencia absoluta de la clase i-ésima y el tamaño muestral, denotada por f_i, recibe el nombre de **frecuencia relativa** de la i-ésima clase.

$$f_i = \frac{n_i}{n}$$

La serie de frecuencias relativas, para las k clases, se indica como
$f_1, f_2, \ldots, f_k$ en el caso de datos categóricos como cuantitativos
Las frecuencias relativas tienen la siguiente propiedad: su suma es igual a la unidad,

$$\sum_{i=1}^{k} f_i = 1$$

Con un sentido práctico suele hablarse de **frecuencias porcentuales**, cuando las fi se las expresa en por ciento, y entonces resulta que su suma es igual al 100%.

Definición 4.18. Las frecuencias absolutas acumuladas, se definen como la frecuencia que resulta de la acumulación, fila por fila, de las correspondientes frecuencias absolutas. La acumulación puede hacerse de dos formas, y según esto resultan:
a**) Frecuencias acumuladas ascendentes**, simbolizadas por $\mathbf{F_i}$: para la i-ésima clase, la frecuencia acumulada ascendente se obtiene sumando a la correspondiente frecuencia, las frecuencias de todas las clases que anteceden a la considerada
$F_1 = n_1$
$F_2 = n_1 + n_2$,
$F_3 = n_1 + n_2 + n_3$, y así sucesivamente hasta la última clase
$F_k = n_1 + n_2 + ... + n_k = \Sigma\, n_i = n$, para $k < n$.
b**) Frecuencias acumuladas descendentes**, simbolizadas por F'_i (que se lee F comilla sub- i): para la i-ésima clase, se obtienen restando a la correspondiente frecuencia, las frecuencias de todas las clases que anteceden a la considerada
$F'_1 = n$
$F'_2 = n - n_1$,
$F'_3 = n - (n_1 + n_2)$ y así sucesivamente hasta la última clase
$F'_k = n - (n_1 + n_2 + ... + n_{k-1})$
Las frecuencias acumuladas ascendentes también son llamadas **frecuencias "menor que**", y las descendentes, **frecuencias "mayor que".** Con un criterio análogo se pueden obtener también las correspondientes frecuencias relativas acumuladas.

A continuación se desarrollará el análisis estadístico descriptivo de las distribuciones de datos cuantitativos agrupados.

4.6. PROPIEDADES DE LA DISTRIBUCIÓN DE VARIABLES CUANTITATIVAS EN MUESTRAS GRANDES

En el caso de muestras pequeñas de variables cuantitativas, se vio que las medidas descriptivas estuvieron referidas a dos propiedades de los colectivos de datos: la posición y la dispersión. En muestras grandes, el hecho de tener una distribución de datos agrupados, lleva a la utilización de un número mayor de propiedades. Las mismas se indicaron en la presentación integrada que se hizo sobre las propiedades estadísticas de las variables cuantitativas.

Propiedades estadísticas a describir en: muestras grandes de datos cuantitativos

Tamaño	Propiedades
Grande	Posición (tendencia central y otra) Dispersión Forma: Asimetría y Curtosis

Estas propiedades se miden objetivamente a través de los estadígrafos correspondientes:

1º) **Medidas de posición:** apuntan a los datos más "típicos" de la distribución, como por ejemplo, los que más se repiten y los que ocupan los lugares centrales.

2º) **Medidas de dispersión:** describen si los datos son homogéneos o sea si se diferencianpoco entre sí (variación pequeña) o, si por el contrario, son heterogéneos o muy dispares (variación grande), y también si el patrón de variación presenta regularidad estadística o no.

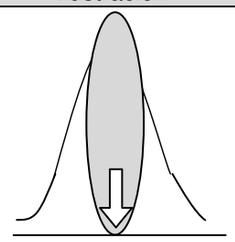
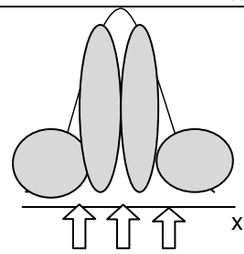
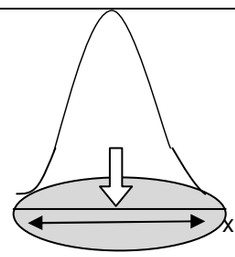
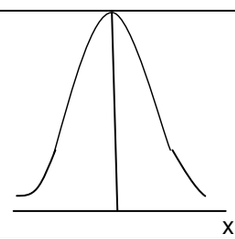
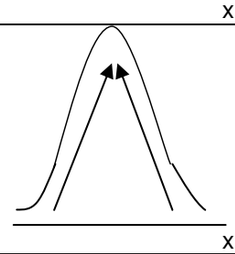
3º) **Medidas de asimetría:** miden en qué grado las distribuciones son asimétricas, a partir de tomar como referencia la media aritmética y considerar si los datos se reparten análogamente a ambos lados de ella. La falta de **simetría** lleva a hablar de **distribuciones sesgadas**.

4º) **Medidas de curtosis:** cuantifican el grado de agudeza o apuntalamiento de la distribución en la parte central, dada por una concentración de los datos (frecuencias más altas) alrededor de la media, y el grado de alejamiento que poseen los valores extremos.

En general estas medidas han sido establecidas procurando que cumplan, lo cual logran en mayor o menor grado, ciertas condiciones entre las cuales se tienen las siguientes:

- Deben tener una definición objetiva, para que distintas personas puedan llegar a partir de un mismo conjunto de datos a un mismo resultado numérico y conclusiones.
- Deben basarse en lo posible en todos los datos de la variable, de forma que la medida no sea inestable, esto es que cambie sustancialmente con sólo variar un valor de variable
- Deben ser fáciles de calcular e interpretar.

Cuadro 4.2: Síntesis de las propiedades estadísticas para muestras grandes de datos cuantitativos

Propiedad		Concepto	Ilustración	Medida
POSICIONAMIENTO	Central (Promedios)	Propensión de los datos (valores de la variable) a ubicarse en el entorno de un punto central de la distribución, correspondiente al recorrido de la variable, donde se ubica el punto de equilibrio.		Estadígrafos de tendencia central, por ej.: la media
	Otro (cuantiles)	Ubicación de puntos en la escala correspondiente al recorrido de la variable (valores de variable), relacionados con la partición de la distribución de datos de modo de dejar en cada una de las partes igual cantidad de datos (comúnmente 1%, 5%, 10%, 25% o el 50%).		Estadígrafos de posición, por ej.: cuartiles (1/4 = 25% en cada parte)
DISPERSIÓN		Grado de fluctuación de los datos, referenciada a un valor central de la variable, de modo aproximado o distante entre sí.		Estadígrafos de dispersión, por ej.: amplitud.
FORMA	Asimetría	Forma de distribución de los datos, a ambos lados de un eje ubicado en el centrado de la distribución.		Estadígrafos de asimetría, por ej.: coeficiente de asimetría.
	Curtosis	Forma de concentrarse los datos, alrededor del centrado de la distribución, que determina un mayor o menor apuntalamiento de la distribución.		Estadígrafos de curtosis, por ej.: coeficiente de curtosis.

4.7. VARIABLES DISCRETAS

Se partirá de un conjunto de datos muestrales, correspondientes a un experimento donde se registró el número de flores por planta, en 50 plantas seleccionadas al azar. Primeramente se procederá a identificar algunos aspectos que definen las características del problema que conducen a la elección del camino a seguir.

Variable observada	Unidad de muestra y análisis	Tipo de dato	Tamaño muestral
Nº de flores/planta	planta	Cuantitativo discreto	n=50

Tabla auxiliar. Registros del recuento de flores (datos de campo)

10	*8*	*6*	*3*	*9*	*7*	*5*	*4*	*6*	*9*
8	*10*	*7*	*9*	*10*	*6*	*8*	*6*	*3*	*2*
4	*3*	*2*	*7*	*5*	*5*	*4*	*3*	*7*	*6*
6	*7*	*8*	*8*	*6*	*7*	*7*	*9*	*8*	*6*
5	*3*	*2*	*1*	*4*	*3*	*6*	*8*	*7*	*0*

4.7.1. Herramientas de análisis gráfico

Presentación tabular

A continuación se presenta la estructura mínima de una tabla de distribución de frecuencias para una variables discreta (tabla modelo). En ella se pueden reconocer: una primera columna que muestra los posibles valores de la variable (x_i, donde i=1,2,…,k) y otra para los datos de frecuencia absolutas (n_i), aunque podrían haberse utilizado las frecuencias relativas o las porcentuales.

Tabla básica de distribución de frecuencias para una variable discreta

x_i	n_i
x_1	n_1
x_2	n_2
.	.
.	.
x_k	n_k
	n

A continuación se muestra la **tabla completa de distribución de frecuencias** que se utilizaría para presentar los resultados del trabajo**.**

Tabla 4.3. Distribución del número de flores por planta

Nº de flores, (1)	Cantidad de plantas (2)	Cantidad de plantas acumulada		Proporción de plantas (5)	Proporción porcentual (6)
		"nº menor o igual que" (3)	"nº mayor o igual que" (4)		
0	1	1	50	0,02	2,0
1	1	2	49	0,02	2,0
2	3	5	48	0,06	6,0
3	6	11	45	0,12	12,0
4	4	15	39	0,08	8,0
5	4	19	35	0,08	8,0
6	9	28	31	0,18	18,0
7	8	36	22	0,16	16,0
8	7	43	14	0,14	14,0
9	4	47	7	0,08	8,0
10	3	50	3	0,06	6,0
	50	--	--	1,00	100,0

Construcción
(1) **valores observados** de la variable.(x_i)
(2) **frecuencia absoluta** (n_i). Notar el total, n= 50
(3) frecuencias acumuladas ascendentes (F_i)
(4) frecuencias acumuladas descendentes (F'_i)
(5) **frecuencias relativas (**f_i). Notar el total, sum (f_i)= 1
(6) **frecuencias relativas porcentuales (**% f_i**).** Notar el total, sum (% f_i)= 100

¿Cuál es la información se puede obtener de la tabla de frecuencias así construida?
Se puede ver que el número total de datos es 50, que las plantas tuvieron entre 0 y 10 flores.
Las plantas con menos de 3 flores y con más de 9 son poco frecuentes, que plantas que tienen entre 6 y 8 flores son las típicas (mayores frecuencias), y que el valor más repetido ha sido 7.
El 18% de las plantas presentaron 6 flores, un 2% fueron plantas sin flores y un 6% (3 plantas) fueron muy floríferas, para ellas se registró un valor máximo de 10 flores
Un 10% de las plantas tuvieron 2 o menos flores, 30% tuvo 4 o menos flores y, casi la mitad de las plantas tuvo entre 0 y 6 flores/planta.

Se deja al alumno, el ejercicio de realizar otras interpretaciones, a partir de la lectura de esta tabla de frecuencias. Realmente extraer esta información a partir de los datos sin procesar, hubiera sido extremadamente dificultosa.

Representaciones gráficas

Gráfica de líneas

Para el ejemplo de variable discreta que se está analizando se tiene lo siguiente:

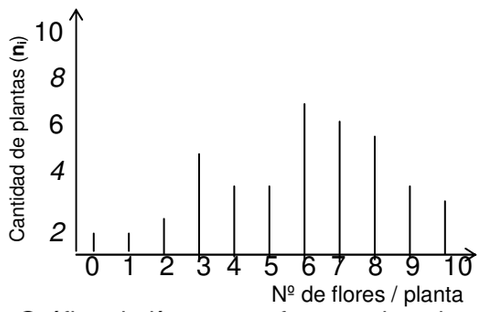


Gráfico de líneas con frecuencias absolutas

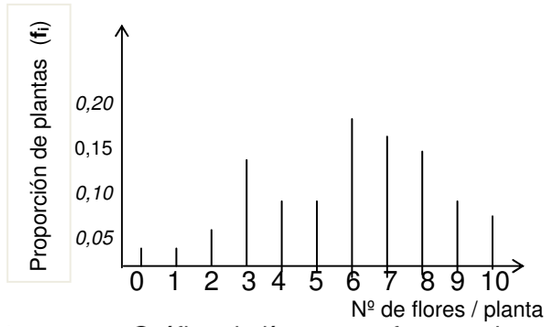


Gráfico de líneas con frecuencias relativas

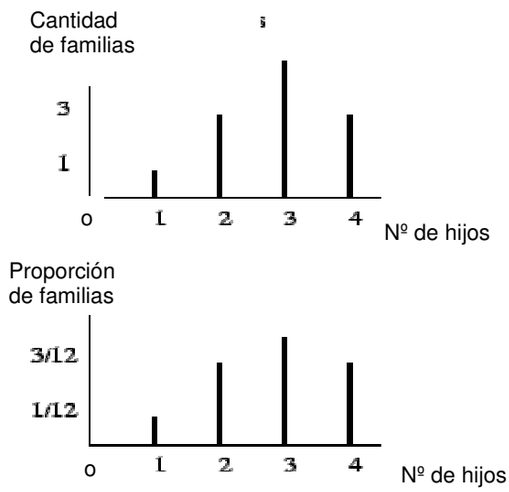
Construcción: Si en el eje de las abscisas se consideran los distintos valores que toma la variable y, en el eje de las ordenadas se consideran las frecuencias absolutas (o las frecuencias relativas) y, por los puntos resultantes se bajan líneas hasta las abscisas, se obtiene un gráfico de líneas para frecuencias absolutas (o de frecuencias relativas).

Gráfica escalonada

Existe también la posibilidad de utilizar representaciones que permitan obtener información de tipo integral, por ejemplo, que permitan encontrar la respuesta al siguiente interrogante ¿cuántas unidades de análisis muestrales presentan un valor igual o menor a un cierto x_i?. Es decir gráficas que se basen en los valores de frecuencias acumuladas, que para el caso de una variable discreta mostrarán un patrón escalonado de frecuencias. Sea por ejemplo, una muestra de datos correspondientes al número de hijos/familia de cierta zona rural y la correspondiente tabla de frecuencias.

Número de hijos (x_i)	1	2	3	4
Cantidad de familias (n_i)	1	3	5	3

Valor de variable	Frec. absolutas	Frec. acum. ascendentes	Frec. relativas
x_i	n_i	F_i	f_i
1	1	1	0,083
2	3	4	0,250
3	5	9	0,416
4	3	12	0,250
Total	12	-	≅ 1,000



Diagramas de barras para frecuencias absolutas y frecuencias relativas.

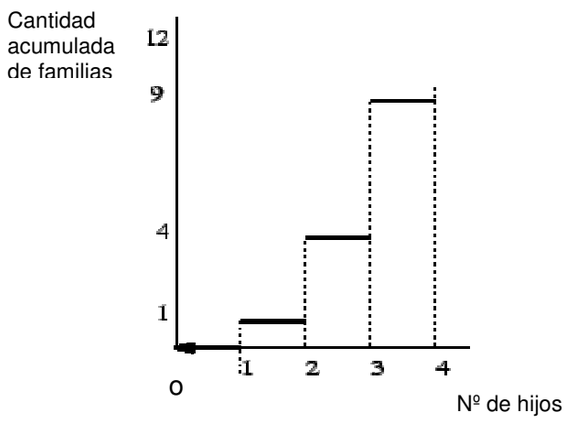


Diagrama de frecuencias acumuladas "menor que" o diagrama escalonado ascendente

Resumen. Gráficos para distribuciones de frecuencias de variables estadísticas cuantitativas discretas

Diagrama de líneas para valores puntuales de la variable observada según su frecuencia. Muestra para cada valor observado (x_i) de la variable, la correspondiente frecuencia de presentación en la muestra.

Eje y, pueden utilizarse

n_i : frecuencias absolutas
f_i : frecuencias relativas
100 f_i : porcentajes

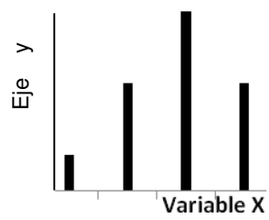


Gráfico (a)

Gráfico escalonado. Distribución de frecuencias acumulados: El gráfico (b) de frecuencias ascendentes muestra al producirse el salto en cada escalón la cantidad de unidades observadas con valores "iguales o menores" al correspondiente x_i. El último escalón (quinto escalón) indica el total de los datos menor o igual al valor máximo observado (x_4), por lo que al mismo tiempo se refiere a todas las unidades medidas (n, o 100%). En forma análoga se puede interpretar un gráfico de frecuencias descendentes.

Eje y, pueden utilizarse

F_i : frecuencias absolutas acumuladas ascendentes
F'_i : frecuencias absolutas acumuladas descendentes

También pueden usarse frecuencias relativas acumuladas y porcentajes acumulados.

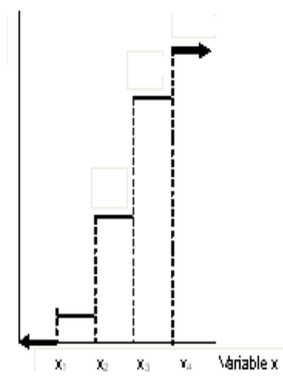


Gráfico (b)

Herramientas de análisis numérico (Estadígrafos)

Medidas de posición

Se analizarán a continuación las medidas de posición, recordando que éstas pueden representar la centralización en torno a la cual se distribuyen la mayoría de las mediciones o bien a otras posiciones. Entre las primeras se tienen aquellas que en general reciben el nombre de promedios (diferentes tipos de medias) y otras como la mediana y la moda. Entre las segundas están medidas que mayoritariamente se refieren a posicionamientos no centrales (cuartiles, deciles y percentiles).

4.7.2.1.1. Media aritmética

En el caso en que los datos estuviesen agrupados en una tabla de Tipo I, es decir, si existen k valores distintos de la variable X, esto es x_1, x_2, ..., x_k, se tienen k clases numéricas, tales que cada valor x_i se repite n_i veces, entonces, la expresión para la media aritmética es:

Definición 4.19. *La media muestral de una variable discreta se calcula como*

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

siendo: x_i *: dato observado,* n_i *la frecuencia absoluta correspondiente de modo que* $n = \sum_{i=1}^{k} n_i$ *y k, el número de valores diferentes que toma la variable observada*

Ejemplo 4.1*:* Si medimos el número de hijos de 15 obreros rurales de una cierta Industria, y los resultados arrojan la siguiente tabla de tipo I, entonces, el número medio de hijos por empleado es:

Nº de hijos (x_i)	Nº de obreros (n_i)	$x_i n_i$
0	2	0
1	4	4
2	8	16
3	5	15
4	1	4
Total	n=20	Σ=39

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n} = (39/20) = 1.95 \quad hijos \cong 2 \quad hijos$$

Debe notarse que en el cálculo de la media intervienen todos los valores de la variable, de ahí que resulte por excelencia la medida promedio que caracteriza el lugar central de la distribución.

4.7.2.1.2. Mediana

A diferencia de la media, la mediana es una medida que trata de caracterizar un posicionamiento que equilibre la cantidad de frecuencias observadas a uno y otro lado. Para encontrar cual es el valor mediano de una distribución de frecuencias discretas, se trabaja con una tabla de frecuencias acumuladas de menor a mayor. La mediana es igual al primer valor de variable, que acumulando las frecuencias, deja por debajo un 50% de las observaciones. En el siguiente ejemplo se observa que la mediana es igual a 2: hay 50 fábricas con valores menores o iguales a ella, y también 50 fábricas con valores igual o mayores a ella.

Ejemplo 4.2: Número de empleados de 100 pequeñas fábricas

Nº de empleados (x_i)	Nº de fábricas (n_i)	F_i
2	20	20
2	30	50
3	25	75
4	15	90
5	10	100
Total	n=100	---

Esta es, como ya vimos, una medida de posición, generalmente central, que se fundamenta en las frecuencias de la distribución. Pero, conviene en este momento tener una visión amplia que aclare cuando corresponde utilizarla, por cuanto, muchas veces se la aplica mal. Para ello hay que tener en cuenta el tamaño muestral y el tipo de variable:
1º) en principio, la moda tiene no tiene sentido en muestras pequeñas y sí, en muestras grandes, porque su valor es muy inestable
2º) a su vez, siendo la muestra grande, la moda tiene sentido en los siguientes casos:

4.7.2.1.3. Moda

La moda es el valor de la variable que más se repite. Cuando la variable es discreta, solo se necesita observar en su distribución de frecuencias cual es el valor de variable que tiene la mayor frecuencia absoluta.
Determinación de la moda

En distribuciones tipo I con clases numéricas: su determinación es inmediata, solo basta observar el valor o valores de la variable que tengan máximas frecuencias con relación a las restantes frecuencias de la distribución.

Evidentemente, cualquier moda absoluta será, pues, una moda relativa. Sin embargo, lo contrario no es en absoluto siempre cierto. Veamos algunos ejemplos ilustrativos, utilizando diferentes distribuciones muestrales de una variable discreta.

Muestra 1	x_i	0	1	2	3	4	5	6
	n_i	7	10	12	25	20	13	5

Muestra 2	x_i	0	1	2	3	4	5	6
	n_i	3	17	12	20	35	10	6

Muestra 3	x_i	0	1	2	3	4	5	6
	n_i	4	15	15	12	28	15	5

Se puede identificar lo siguiente:

Muestra 1: se destaca una sola frecuencia, la cual es igual a 25, por tanto se tiene una moda absoluta igual a 3,
Muestra 2: se tienen dos frecuencias que llaman la atención, 17 y 35, los valores correspondientes de variable 1 y 4 son modas relativas, y además 4 es una moda absoluta (distribución bimodal),
Muestra 3, se tienen tres modas relativas que son 1,2 y 4, en correspondencia con las frecuencias destacadas en la serie 15 y 18, pero sólo 4 es moda absoluta (distribución trimodal). Nótese que el valor 5, asociado a una frecuencia igual a 15, no es moda porque no se destaca entre los valores contiguos.

***Ejemplo 4.3*:** Sea el número de salames con principio de enranciamiento en ristras de tamaño 5, seleccionadas aleatoriamente de estanterías comerciales de almacenes mayoristas.

Número de salamines rancios, x_i	0	1	2	3	4	5
Cantidad de salamines, n_i	5	18	18	9	3	2

Se observa que la distribución tiene dos modas relativas, ya que la máxima frecuencia, igual a 18, corresponde tanto al valor de variable 1 como 2.

Medidas de dispersión

4.7.2.2.1. Amplitud muestral

También se denomina rango o recorrido. Es válido lo visto para muestras pequeñas.

4.7.2.2.2. Varianza muestral

En el caso de variables discretas, se tienen k diferentes valores x_i, La fórmula (a) se basa en los cuadrados de *k desvíos respecto a la media* $\left(x_i - \bar{x}\right)$, mientras que la fórmula (b) se basa en los *k valores observados de la variable* x_i

(a) Procedimiento directo

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1} \quad ; \textit{siendo} \quad i = 1,2,\ldots,k$$

Fórmula definicional: suma de cuadrados de desvíos ponderados por las frecuencias absolutas, dividida por los grados de libertad.

(b) ***Procedimiento abreviado***

$$s^2 = \frac{\sum_{i=1}^{k} x_i^{\,2}\, n_i - \frac{\left(\sum_{i=1}^{k} x_i . n_i\right)^2}{n}}{n-1} \quad ; \; i = 1,2,\ldots,k$$

Notar:

$\sum_{i=1}^{k} x_i^2\, n_i$ ***suma de k productos****, y se ponderan los* ***cuadrados de los valores observados de la variable*** x_i^2

$\left(\sum_{i=1}^{k} x_i\, n_i\right)^2$ ***cuadrado de la suma de k productos****, y se ponderan los* ***valores observados de la variable****:* $\boldsymbol{x_i}$

4.7.2.2.3. Desviación típica muestral

La **desviación típica**, se obtiene según ya se ha visto como la raíz cuadrada positiva de la varianza

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

4.7.2.2.4 Coeficiente de variación muestral

Es válido lo visto para muestras pequeñas: $cv = \frac{s}{\bar{x}}$ o bien $\% \, cv = 100 \left(\frac{s}{\bar{x}} \right)$

Medidas de forma: asimetría y curtosis

Estas medidas serán desarrolladas en forma integrada para el caso de las variables discretas y continuas, después de presentar el análisis descriptivo de las variables continuas.

4.8. VARIABLE CONTINUA

4.8.1. Herramientas de análisis gráfico

4.8.1.1. Presentación tabular

Para describir la distribución de frecuencia correspondiente a una variable continua, es indispensable agrupar los valores registrados mediante un conjunto de intervalos de clase.

Ejemplo 4.4: La siguiente es la tabla primaria correspondiente a un estudio sobre el perímetro, en centímetros, a la altura de la primera ramificación, de troncos de damasco variedad Royal, de un monte frutal de 4 años, realizado en Lavalle en 1974.

35	38	48	45	43	28	42	54	41	49	36	58
48	48	38	45	31	29	54	58	46	36	26	46
42	42	39	39	42	34	28	55	36	40	49	52
42	49	40	35	53	31	29	28	35	54	49	36
36	43	33	43	46	32	40	36	41	36	33	40
38	58	40	45	45	49	45	33	34	46	36	43
42	39	30	31	49	45	55	25	37	45	34	53
44	48	40	30	42	56	44	27	48	44	41	40
39	45	33	41	34	27	50	24	46	43	45	36
43	43	42	35	33	52	48	37	42	40	49	41

Trabajando como se vio en el Tema 2, para esta muestra se tiene lo siguiente:

1º) Amplitud muestral, a partir de los límites reales de la muestra,
$\Delta m = x_{max} - x_{min} = 58 \text{ cm} - 24 \text{ cm} = 34 \text{ cm}$

2º) Número de intervalos de clase, utilizando la fórmula de Sturges

$$k = 1 + 3{,}3 \cdot \log 120 = 7{,}86$$

En principio, el nº de intervalos que debería usarse en este caso sería 8. Sin embargo, recordemos que es aconsejable que este número sea impar, en consecuencia podría decidirse usar 7 ó 9 intervalos. Se optará por el primer número porque el tamaño muestral no es grande y además porque 7 se aproxima más al valor calculado según la fórmula.

3º) Longitud de los intervalos de clase
$\Delta x = \Delta m / k = 34 \text{ cm} / 7 \cong 5 \text{ cm}$

4º) Clasificación de los datos

Tabla 4.4. Tabla auxiliar para la clasificación de los datos

Intervalo de clase **discreto**	Clasificación del dato	Número de troncos (n_i)
(25-29]	~~////~~ ///	8
(30-34]	~~////~~ ~~////~~ ///	13
(35-39]	//// //// //// //// /	21
(40-44]	//// //// //// //// //// //// ///	33
(45-49]	//// //// //// //// ///	23
(50-54]	~~////~~ ~~////~~ ///	13
(55-59]	~~////~~ ////	9

Definición 4.20.
El valor promedio entre los límites del intervalo se llama punto medio del intervalo o "marca de clase". Este valor es un promedio que se usa para representar a todos los datos que se clasificaron en el intervalo, por lo tanto, constituye un valor de variable no observado, pero muy útil para realizar los cálculos posteriores. Como es un valor de variable, se lo denota con "x_i".

La distribución de frecuencia se puede presentar en una tabla básica, donde los intervalos se ponen en correspondencia con las frecuencias absolutas. Sin embargo, para mejorar el análisis, casi siempre es deseable elaborar la distribución de frecuencia relativa o la distribución porcentual, dependiendo de si se prefieren las proporciones o los porcentajes.

Tabla de distribución de frecuencias completa

Tabla 4.5. Distribución de frecuencias de perímetros de troncos de damascos (en cm), variedad Royal, de 4 años. Lavalle, 1994.

Intervalo de clase **continuo**	Punto medio	Frec. Absoluta	Frec. Acumulada.		Frec. relativa	Frec. relativa acumulada
			Ascen.	Desc.		
24,5 –29,5	27,0	8	8	120	0,067	0,067
29,5 –34,5	32,0	13	21	112	0,108	0,175
34,5 –39,5	37,0	21	42	99	0,175	0,350
39,5 –44,5	42,0	33	75	78	0,275	0,625
44,5 –49,5	47,0	23	98	45	0,192	0,817
49,5 –54,5	52,0	13	111	22	0,108	0,925
54,5 -59,5	57,0	9	120	9	0,075	1,000
-	-	120	-	-	1,000	-

Tabla de distribución porcentual

Como se anticipó, la utilidad de la distribución de frecuencia relativa o de la distribución porcentual es grande cuando se comparan muestras diferentes, especialmente si el tamaño muestral no es igual. Se emplean los valores de las frecuencias relativas multiplicados por 100, de modo parcial (Tabla 4.6) o bien acumuladas.

Ejemplo 4.5: A partir de los datos del censo nacional agropecuario se ha analizado la distribución la cantidad de hectáreas incultas por finca en una cierta zona, con el siguiente resultado:

Tabla 4.6. Distribución porcentual de las hectáreas incultas por finca en cierta zona (n=240).

Hectáreas incultas/finca (n=240)	Porcentaje de fincas
10,5 a menos de 20,5	48,9
20,5 a menos de 30,5	26,7
30,5 a menos de 40,5	12,8
40,5 a menos de 50,5	6,4
50,5 a menos de 60,5	3,0
60,5 a menos de 70,5	1,5
70,5 a menos de 80,5	0,7
Total	100,0

Tabla 4.7. Distribución porcentual acumulada de las hectáreas incultas por finca, menor al valor dado (n=240)

Hectáreas incultas/ finca	Porcentaje de fincas "menor que"
<20,5	48,9
<30,5	75,6
<40,5	88,4
<50,5	94,8
<60,5	97,8
<70,5	99,3
<80,5	100,0

Ref.: el valor mínimo de la variable fue 10,5 hectáreas

Interpretaciones:

- La tercera fila en la Tabla 4.6 indica que un 12,8 % de las 240 fincas poseen una superficie inculta mayor o igual a 30,5 hectáreas y no mayor a 40,5
- La tercera fila en la Tabla 4.7 indica que hay un 75,6% de fincas con una superficie inculta menor a 30,5 hectáreas.

En forma análoga, se puede construir una tabla que muestre la distribución porcentual acumulada mayor que el límite inferior de la variable.

Tabla 4.8. Distribución porcentual acumulada de las hectáreas incultas por finca, mayor al valor dado (n=240)

Límite inferior	Porcentaje de fincas "mayor que"
>10,5	100,0
>20.5	51,1
>30.5	24,4
>40,5	11,6
>50,5	5,2
>60,5	2,2
>70,5	0,7
>80.5	0.0

Una importante observación

En las tablas puede observarse que la **frecuencia relativa** tiene dos aspectos de gran interés:
1º) expresada en % resulta muy fácil de interpretar el significado y además facilita la comparación entre muestras que tienen diferente tamaño.
2º) desde un punto de vista más teórico, se la puede considerar como una estimación empírica de la probabilidad de ocurrencia de algún suceso empírico. Por tratarse de proporciones, una propiedad que cumplen las frecuencias relativas es que sus valores varían en el intervalo [0 ; 1] y, del mismo la función probabilidad que se estudiará en la Unidad de Probabilidad se define numéricamente en un intervalo [0 ; 1]. En el caso de las probabilidades, el 0 indica que un suceso es imposible (por ejemplo, que al tirar un dado de seis caras, resulte una cara con siete puntos) en tanto que el 1 indica que el suceso va a ocurrir con certeza (por ejemplo, que al tirar un dado de seis caras, resulte una cara con 1 a 6 puntos). En la realidad cuanto más probable es que ocurra un suceso, por lo general la frecuencia relativa correspondiente a lo observado resultará más cercana a 1, y cuanto menos probable sea su ocurrencia, por lo general la frecuencia relativa correspondiente a lo observado resultará más cercana a 0.

La frecuencia relativa, permite intuir algunas fundamentales propiedades de la probabilidad.

4.8.1.2.Representación gráfica

Histograma

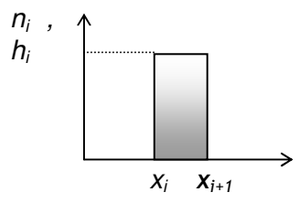
Definición 4.21..
Un **histograma** consiste en una serie de rectángulos adyacentes (en el diagrama de barras son no adyacentes), cuyo ancho es proporcional al alcance de los datos que se encuentran dentro de una clase, y cuya altura es proporcional al número de elementos que caen dentro de la clase.

Si las clases que utilizamos en la distribución de frecuencias son del mismo ancho, lo más común, entonces que las barras verticales del histograma también tengan el mismo ancho. La altura de la barra correspondiente a cada clase representa el número de observaciones de la clase o frecuencia. Como consecuencia de lo anterior, el área de cada barra del histograma puede ser:

Proporcional a la frecuencia de clase, si en ordenadas se representan las frecuencias (n_i)

$A = b \cdot h$

$A = \Delta x \cdot n_i$



Igual a la frecuencia de clase, si en ordenadas se representa la altura o densidad de clase (h_i), que es $x_i / \Delta x$.

$A = \Delta x \cdot h_i$; $h_i = n_i / \Delta x$
$A = \Delta x \, (n_i / \Delta x)$
$A = n_i$

Un histograma que utiliza las frecuencias relativas de los puntos de datos de cada una de las clases, en lugar de usar el número de puntos, se conoce como **histograma de frecuencias relativas**. Este tipo de histograma tiene la misma forma que un histograma de frecuencias absolutas construido a partir del mismo conjunto de datos. Esto es así debido a que en ambos, el tamaño relativo de cada rectángulo es la frecuencia de esa clase comparada con el número total de observaciones.

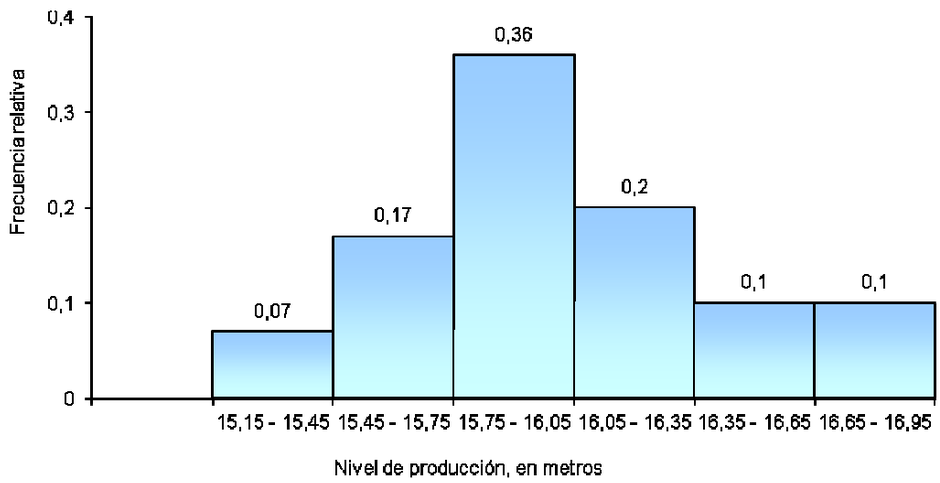


Gráfico 4.1. Distribución de frecuencias relativas de los niveles de producción, en metros.

Ventajas de un histograma de frecuencias relativas:

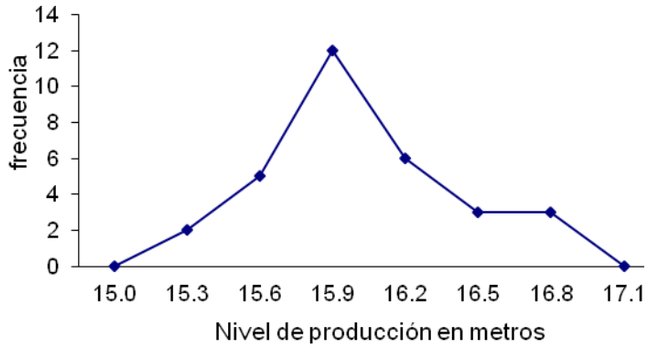
Presentar los datos en términos de la frecuencia relativa de las observaciones, más que en términos de la frecuencia absoluta, es de utilidad ya que mientras los números absolutos pueden sufrir cambios, la relación entre las clases permanece estable.

Resulta fácil comparar los datos de muestras de diferentes tamaños cuando utilizamos histogramas de frecuencias relativas. Sin embargo, cuando se comparan dos o más conjuntos de datos, no es posible construir los diversos histogramas en la misma gráfica, porque la superposición de barras verticales dificulta su interpretación. Para ese caso, es necesario construir polígonos porcentuales o de frecuencia relativa.

Polígono de frecuencias

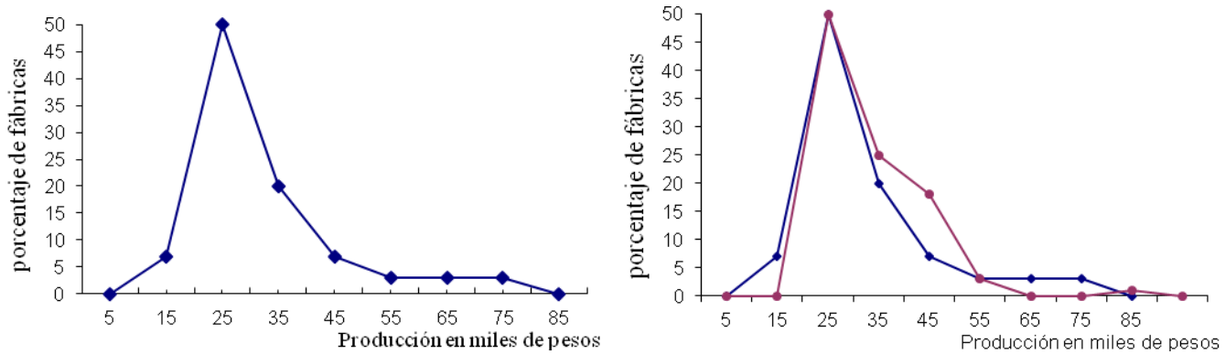
Los polígonos de frecuencias son otra forma de representar gráficamente distribuciones, tanto de frecuencias simples como relativas.

Construcción. Para construir un **polígono de frecuencias**, en el eje de abscisas señalamos, como en el histograma, los valores de la variable pero en este caso corresponde usar los puntos medios. A continuación, graficamos los puntos en correspondencia a las frecuencias de clase (proyectando por sobre el valor del punto medio) y conectamos los puntos resultantes sucesivos con segmentos, de modo que resulta una línea irregular (quebrada) abierta. Finalmente se cierran los extremos (límite inferior y límite superior) formando un polígono (una figura con muchos lados).



Si se compara la figura que representa un polígono de frecuencias con el gráfico del histograma anterior, se dará cuenta que se han añadido dos clases, una en cada extremo de la escala de valores observados. Estas dos nuevas clases contienen cero observaciones, pero permiten que el polígono alcance el eje horizontal en ambos extremos de la distribución (100% área).

El **polígono porcentual** se forma haciendo que el punto medio de cada clase represente los datos de esa clase y después conectando la secuencia de sus respectivos porcentajes de clase.



Polígonos de frecuencia porcentual

Construcción de un **polígono de frecuencias relativas**: Un polígono de frecuencias que utiliza frecuencias relativas de puntos de datos en cada una de las clases, en lugar del número real de puntos, se conoce como polígono de frecuencias relativas. Este polígono tiene la misma forma que el polígono de frecuencias construido a partir del mismo conjunto de datos, pero con una escala diferente en los valores del eje vertical. Más que el número absoluto de observaciones, la escala es el número de observaciones de cada clase como una fracción del número total de observaciones.

Análisis comparativo de ventajas

Histograma	Polígonos de frecuencias
Los rectángulos muestran cada clase de la distribución por separado. El área de cada rectángulo, en relación con el resto, muestra la proporción del número total de observaciones que se encuentran en esa clase.	El polígono de frecuencia es más sencillo que su correspondiente histograma. Traza con más claridad el perfil del patrón de los datos. El polígono se vuelve cada vez más liso y parecido a una curva conforme aumentamos el número de clases y el número de observaciones.

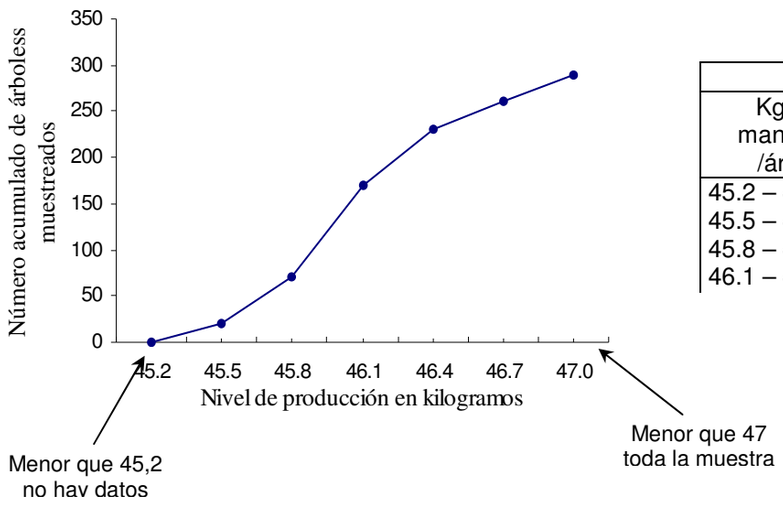
Polígonos de frecuencias acumuladas u ojivas.

Una distribución de frecuencias acumuladas nos permite ver cuántas observaciones están por encima, o por debajo, de ciertos valores.

Polígono de frecuencias acumuladas "menor que" u ojiva ascendente: Los puntos representados en la gráfica indican la cantidad de datos que tienen un valor de variable igual o menor que el valor correspondiente al límite superior del intervalo de clase (eje de abscisas). Observar lo siguiente: el polígono comienza con ordenada cero en el límite superior de un intervalo imaginario anterior (coincide con el inferior del primer intervalo de clase para los valores observados) y termina con ordenada igual a n, en el límite superior de la última clase.

Polígono de frecuencias acumuladas "mayor que" u ojiva descendente: Los puntos representados en la gráfica indican la cantidad de datos que tienen un valor de variable igual o maqyor que el valor

correspondiente al límite inferior del intervalo de clase (eje de abscisas). En este caso el polígono comienza con ordenada igual a n en coincidencia con el límite inferior de un intervalo imaginario anterior (coincide con el inferior del primer intervalo de clase para los valores observados) y termina con ordenada igual a n, en el límite superior de la última clase.



Clasificación	
Kg de manzana /árbol	Cantidad de árboles
45.2 – 45.4	20
45.5 – 45.7	50
45.8 – 46.0	110
46.1 – 46.3	60

Distribución de niveles de producción “menor que” de una muestra de árboles de manzana.

En forma análoga podría construirse un polígono de frecuencias relativas acumuladas “mayor que”.

Gráficos para distribuciones de frecuencias de variables estadísticas cuantitativas continuas

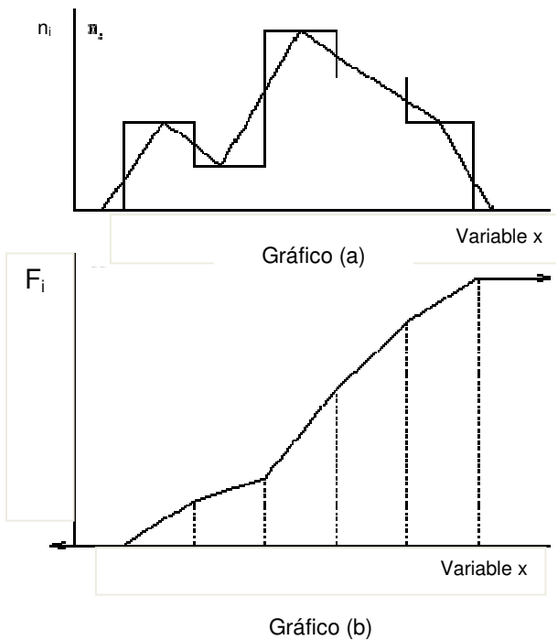


Gráfico (a)
Muestra superpuesta, la silueta del **histograma** con el **polígono de frecuencias**. Notar,
1) que las frecuencias corresponden respectivamente a los intervalos de clase y a los puntos medios, y
2) los puntos de cierre del polígono.

Gráfico (b)

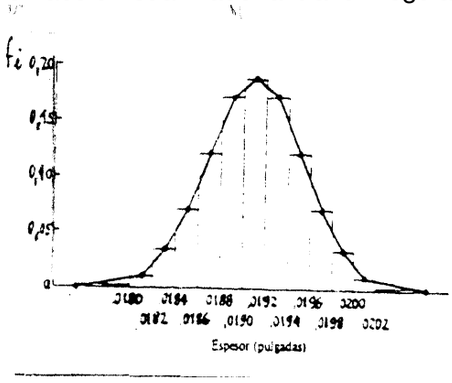
Polígono de frecuencias acumuladas “menor que”, con límites superiores del intervalo (**ojiva ascendente**)

Tablas versus gráficos de distribuciones de frecuencias

Las tablas proporcionan datos numéricos más exactos, mientras que los gráficos solo permiten una lectura aproximada.

La interpretación de tablas con abundantes datos numéricos suele resultar compleja y requiere una buena preparación, en tanto que las representaciones gráficas suelen permitir tomar una idea rápida del fenómeno en estudio. Por ejemplo, la gráfica de una distribución de frecuencias pone en evidencia los patrones de comportamiento de los datos muestrales, con mayor facilidad que las correspondientes tablas.

Las gráficas de variables continuas permiten tomar rápidamente idea acerca del patrón de la distribución poblacional (dado que para ésta se tienen infinitos valores de variable, se tendrían infinitos intervalos de clases, k→∞, y entonces Δx →0). Esto se hace creando una **curva de frecuencias**, **f(x)**, para lo cual se procede a elaborar un polígono de frecuencias relativas, y luego se le hace un suavizado al trazo irregular del polígono.



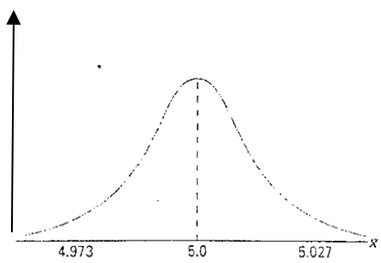


Figura II.33. Área para el ejemplo

4.8.2. Herramientas de análisis numérico: Estadígrafos

Medidas de tendencia central y otras

Media aritmética

En esta situación, siempre haremos la suposición de que, en cada intervalo de la tabla la frecuencia que le corresponde, se encontrará repartida de forma uniforme a lo largo del intervalo, lo que, como consecuencia, da lugar a que el valor medio de cada intervalo coincida exactamente con el punto medio del mismo, y que hemos denominado en un capítulo anterior “marca de la clase” o del intervalo correspondiente, o bien “punto medio”.

Bajo esta hipótesis, la suma del conjunto de valores de un intervalo dado será, pues, igual al producto de su frecuencia por el valor de su marca de clase, sin más que tener en cuenta la interpretación de la media aritmética para los puntos de tal intervalo.

Así, pues, cuando la tabla de datos es de Tipo II y los datos están repartidos entre k intervalos contiguos, cuyas marcas de clase y frecuencias asociadas son, respectivamente, x_i y n_i, la media puede ser obtenida por el siguiente procedimiento.

Definición 4.22
*La **media en distribuciones Tipo II**, es igual a:*

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

Siendo: $n = \sum_{i=1}^{k} n_i$ *y* x_i *: punto medio del i-ésimo intervalo*

Es de notar que, en este caso, para poder disponer de la marca de clase de cada intervalo, se requiere que los intervalos están perfectamente, determinados por unos extremos concretamente definidos. Así, pues, no podríamos calcular la media de una distribución de datos que nos midiera el número de habitantes de los municipios de una provincia, si el grupo de municipios más poblados estuviese definido ambiguamente, diciendo sólo, por ejemplo, que tiene más de 200.000 habitantes.

Ejemplo 4.5: Supongamos que estudiamos el salario anual de los empleados de una fábrica de automóviles y tenemos los datos de dichos salarios recogidos en la siguiente tabla de tipo II:

Miles de Pesos	Marcas de clase (x_i)	Nº empleados	$x_i\, n_i$
50,5 – 60,5	55,0	12	660
60,5 – 70,5	65,0	20	1300
70,5 – 80,5	75,0	18	1350
80,5 – 100,5	90,0	15	1350
100,5 – 120,5	130,0	5	650
		n=70	5310

$$\bar{x} = \frac{5310}{70} = 75{,}857 \text{ miles de pesos}$$

Precaución: En adelante nos referiremos de forma general con x_i al valor i-ésimo de la variable, pero hay que tener en claro que: a) si se trabaja con una distribución simple o con una distribución Tipo I con clases numéricas, x_i corresponde a un valor medido y, b) si se trabaja con datos de una distribución tipo II, x_i corresponde a la marca de clase o punto medio del intervalo i-ésimo. Con esta notación, la formulación matemática de las medidas puede parecer la misma, pero el significado puede llegar a ser muy diferente

Ventajas y desventajas de la media

Ventajas	Inconvenientes
- Es sencilla de calcular -Está perfectamente definida de forma objetiva, y es única -Tiene un claro significado interpretativo -Para su cálculo se utilizan todos los valores de su distribución	-Los valores extremos muy dispares influyen de forma notable en su valor, haciéndola menos representativa.

A pesar de este inconveniente, por sus ventajas, se puede decir que es la medida de posición central más utilizada.

Existe una variante importante de la media aritmética, de aplicación en aquellas circunstancias en las que se conoce que los valores de la variable no tienen todos la misma importancia para su tratamiento, sino que, por el contrario, existen observaciones que deben ser consideradas como más representativas que otras. A esta variante de la media aritmética se la llama **Media aritmética ponderada.** Para su cálculo se le asocia a cada valor de x_i un peso w_i , que nos medirá su grado de importancia o representatividad dentro de la distribución. Estos pesos w_i serán valores positivos que representarán el número de veces que sus correspondientes valores x_i son más representativos que un valor que tuviese peso asociado a la unidad.

Definición 4.23
*La **media aritmética ponderada** de una distribución de valores $x_1, x_2, \ldots, x_k$ cuyos pesos o importancias relativas $w_1, w_2, \ldots w_k$ respectivamente, se define como*

$$\bar{x}_w = \frac{\sum_{i=1}^{k} x_i \cdot w_i}{\sum_{i=1}^{k} w_i}$$

Obsérvese que la media aritmética ponderada puede considerarse como una media aritmética de una distribución hipotética con los mismos valores que la real, pero en lo que un peso w_i de un valor x_i correspondería a que ese valor x_i se repitiese w_i veces y, por lo tanto, pesase w_i veces más que un valor que sólo apareciese una vez. Tal distribución hipotética estaría, entonces caracterizada por valores x_1, x_2……, x_k con pesos o importancias w_1, w_2,……., w_k respectivamente.

Sin embargo, aunque para comprender intuitivamente el significado de la media aritmética ponderada este razonamiento es válido y es por otra parte, importante remarcar que en él nos hemos referido al caso particular en que los pesos w_i eran números enteros, mientras que en general, dichos pesos pueden ser números reales positivos cualesquiera.

Ejemplo 4.6: Sea el caso de un vino que durante su añejamiento aumenta las cantidades de taninos se tiene una partida de vinos de distintos años, de modo que se pueden otorgar las siguientes importancias relativas.

Tiempo	g/l	Ponderación
Cantidad de taninos a los 6 meses	0,7	1
Cantidad de taninos a los 12 meses	0,7	1
Cantidad de taninos a los 15 meses	1	2
Cantidad de taninos a los 2 años	3	5

Como observamos en la tabla, hemos asignado a los vinos una misma importancia básica de 1 hasta el año, y una importancia 5 veces mayor a los dos años. Bajo estos supuestos, si se quiere sacar un valor promedio de la cantidad de tanino para una muestra de esas partidas de vino, sería:

$$\overline{x}_w = \frac{0,7\times1+0,7\times1+1\times2+3\times5}{1+1+2+5}$$

Mediana

Cuando la distribución se presenta en forma de **tabla de tipo II,** puesto que para este tipo de tablas se asume que la variable evoluciona de una forma continua y uniforme, entonces tendremos que encontrar el valor de la variable al que correspondería la frecuencia n/2. Ahora bien, dicho valor se encontrará en el primer intervalo en que su frecuencia absoluta acumulada sea igual o supere a n/2. Llamemos $l_{i\ (q2)}$ al límite inferior de tal intervalo, al que llamaremos intervalo mediano, y por lo tanto que se lee:

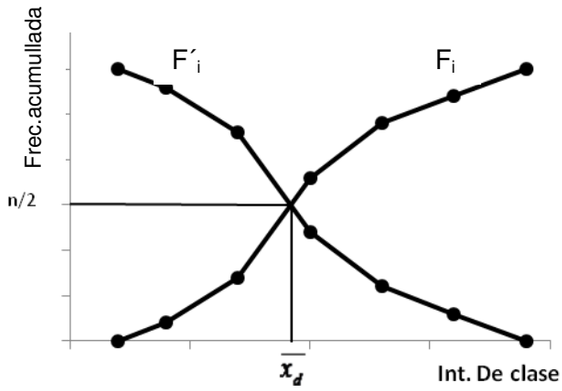
Definición 4.24
La **mediana**, en una distribución de tipo II, es igual al límite inferior del intervalo mediano $l_{i\ (q2)}$ más el cociente que resulta de dividir el valor n/2 menos la frecuencia acumulada hasta el intervalo de clase anterior al mediano $F_{(q2\ -\ 1)}$, por la frecuencia absoluta del intervalo mediano, $n_{(q2)}$, multiplicado por la longitud del intervalo de clase Δ_x.

$$\overline{x}_d = l_{i(q2)} + \frac{n/2 - F_{(q_{2-1})}}{n_{(q2)}} \times \Delta_x$$

Ventajas e inconvenientes de la mediana

Ventajas	Inconvenientes
Es sencilla de calcular Es de fácil interpretación al ser siempre un valor propio de la variable No influye en ella más que los datos centrales de la distribución por lo que se puede calcular aún desconociendo los valores extremos de la distribución, siempre que tengamos suficiente información acerca de sus frecuencias.	No puede expresarse mediante una fórmula matemática sencilla que permita realizar grandes desarrollos algebraicos con ella No intervienen en su confección todos los valores de la variable, sino sólo los centrales. a pesar de todo, este último inconveniente lo es realmente cuando todos los valores de la distribución son conocidos, cosa que no siempre ocurre, y es precisamente en estos casos donde este "inconveniente" se traduce a la tercera "gran ventaja" de la mediana.

La determinación gráfica puede hacerse rápidamente utilizando el polígono de frecuencias acumuladas, y teniendo en cuenta la definición de mediana. La ordenada máxima en este gráfico representa la frecuencia total, o sea n. Dado que la mediana se relaciona con la mitad de los individuos, se individualiza el valor correspondiente a n/2 en el eje vertical. A partir de ese valor se prolonga una línea paralela al eje de abscisas hasta intersectar el polígono de frecuencias acumuladas en el punto A. Desde el punto A luego se baja una perpendicular hasta el eje de abscisas, donde se puede leer el valor de la mediana.

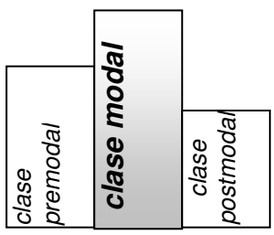


Moda.

Cuando los datos están sin agrupar, se puede hablar de la moda en relación al dato observado con mayor frecuencia, pero cuando los datos están agrupados sólo se puede hablar del intervalo con mayor frecuencia o intervalo modal. Una vez que los datos se han clasificado no es correcto hablar de la moda porque el valor encontrado será teórico, y teóricamente la población es infinita (N→ ∞), en otras palabras la variable toma en cada en cada elemento un valor diferente. Para la variable continua, como veremos en la Unidad de probabilidad, la probabilidad de ocurrencia de un determinado valor es igual a cero, por tanto, hablar de que un valor de variable continua es la moda (tiene la más alta frecuencia) resulta una seria contradicción. Sin embargo, esto no es reflejado por los datos muestrales, debido a que la medición tiene error y entonces aparecen datos repetidos.

Determinación de la moda
Se identifica el (o los) **intervalo modal** donde se clasificó el mayor número de datos y podemos referirnos al **punto medio de la clase modal**, como el valor alrededor del cual se tiene el mayor agrupamiento o densidad de datos.

En el caso de variable continua, también puede hablarse de un intervalo premodal y uno posmodal, como se muestra en el siguiente diagrama:



Definición 4.25
Se llama **moda absoluta**, representada por $\bar{x}_m$, a aquel valor de la variable cuya frecuencia absoluta no es superada por ningún otro valor de la variable en la muestra.

Definición 4.26
Se llama **moda relativa** a aquel valor de la variable cuya frecuencia absoluta asociada no es superada por las de sus valores contiguos.

Ventajas e inconvenientes de la moda

Ventajas	Inconvenientes
Es sencilla de calcular lo modal. En variables discretas es de fácil interpretación, al ser siempre un valor propio de la variable.	No puede expresarse de forma sencilla mediante fórmula matemática que permita operar cómodamente con ella. No detecta ningún cambio en la distribución que se produzca ajeno al valor modal o intervalo modal.

Resulta adecuada una visión integral de las tres medidas descriptas, media, mediana y moda, pero la postergaremos hasta tratar el tema de simetría y sesgo de una distribución.

Cuantiles o fractiles

Las medidas que vamos a ver ahora se llaman medidas de posición no central, porque, aún tratándose de posicionar sobre la escala de posibles valores de la variable algún punto característico de la distribución, ese punto de interés generalmente no es el central. La combinación de estas medidas de posición no necesariamente centrales, con las medidas de posición central, nos permitirá evaluar el

comportamiento de la distribución de frecuencias desde un punto de vista general, a lo largo de todos los valores de la variable, y no concentrándonos en unos pocos de ellos que dicen mucho sobre la tendencia central pero nada acerca de las colas de la distribución, esto es, los valores que se posicionan por debajo de los centrales y por encima de los centrales).

La idea es análoga a la que nos permitió definir la mediana, que, recordemos es un valor de la variable que deja a cada uno de sus lados igual cantidad de datos muestrales (50% por debajo y 50% por encima). Ahora, siempre con los datos de la variable ordenados en forma creciente, nos interesa encontrar cuál de los x_i, deja a su izquierda (incluyéndolo a él) cierta proporción generalmente diferente al 50% de la distribución. Un gráfico dará luz a este nuevo concepto. En (a) se indica el cuantil que deja por debajo (incluyéndolo a él) un 20% de los valores de la variable X y, por encima (incluyéndolo a él) un 80%, mientras que en (b) se da la situación inversa.

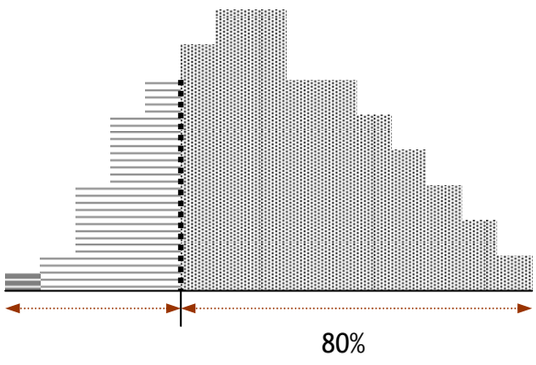


Gráfico (a)

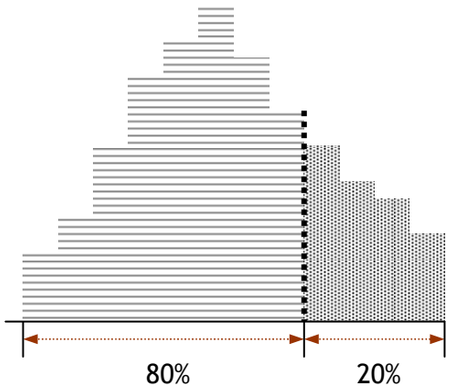


Gráfico (b)

Los **cuantiles** se pueden clasificar en cuatro clases de medidas, de las cuales en este curso, nos interesa en especial la primera y la última:

Cuartiles: dividen la distribución en cuatro partes de igual frecuencia (n/4) , lo que significa que cada parte contiene ¼ del total de datos, es decir, un 25%.
Quintiles: dividen la distribución en cinco partes de igual frecuencia (n/5), lo que significa que cada parte contiene un 20% del total de datos.
Deciles: dividen la distribución en diez partes de igual frecuencia (n/10), lo que significa que cada parte contiene un 10% del total de datos.
Percentiles: dividen la distribución en cien partes de igual frecuencia (n/100), lo que significa que cada parte contiene un 1% del total de datos.

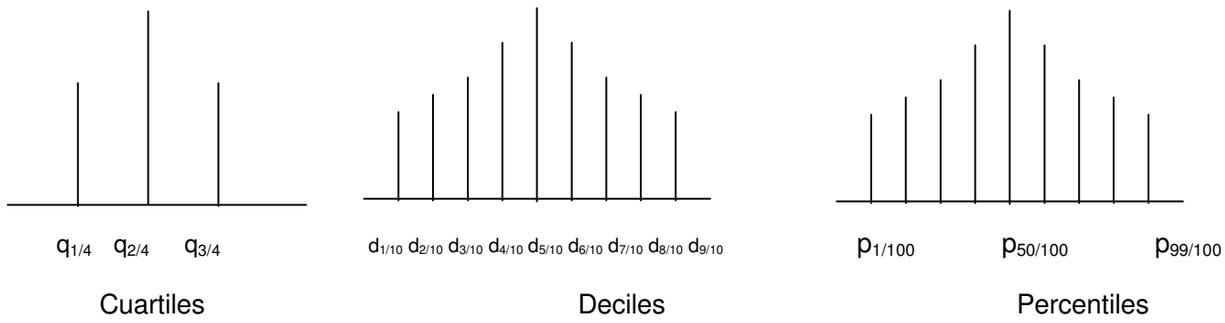
Notar que si los cuantiles dividen en k partes, la cantidad de cuantiles es igual a k-1.

Definición 4.27
Un **cuantil**, que se representa por $\mathbf{q_{r/c}}$ y se lee como "cuantil r-ésimo de orden c", es aquel valor de la variable x_i, que en un arreglo de datos ordenados en forma creciente, permite dividir a la distribución del total de los datos dejando por debajo al menos r/c partes de datos, y por encima al menos las (r/c)/c partes restantes.

Por ejemplo: sea el segundo cuantil de orden 4, esto es $\mathbf{q_{2/4}}$. Primeramente entendemos que nos estamos refiriendo a una distribución dividida en 4 partes (cuartos o cuartiles), y un valor de variable que deja por debajo 2 de esas 4 partes, es decir, la mitad de los datos y por encima el resto, que son otras 2 de esas 4 partes porque 1- (2/4) = 2/4. En otras palabras, nos estamos refiriendo a aquél valor de variable por x_i, que en un arreglo ordenado de menor a mayor, permite dividir la distribución de frecuencias dejando por debajo al menos la mitad de los datos de la distribución, y por encima al menos la otra mitad, o sea, que en definitiva al segundo cuartil, en símbolo $\mathbf{q_{2/4}}$, que es la mediana ya conocida por nosotros.

En forma análoga a la dada para la mediana, se pueden desarrollar fórmulas para el cálculo del primer y tercer cuartil.

Los percentiles serán muy utilizados en inferencia estadística en relación a conceptos probabilísticos. En este contexto, las poblaciones de variables continuas se representan con curvas que se definen mediante funciones f(x), una de las cuales es la curva normal o curva campanular. La función de la normal, es de gran utilidad porque representa a la distribución teórica de muchas variables continuas de interés en Agronomía y Bromatología, y ya resulta familiar a quienes han estudiado la teoría de errores en Física. A partir de ella, mostraremos los gráficos que indican los cuartiles, deciles y percentiles:



Ejemplo 4.7: Sea la variable peso de racimos de uva en gramos. Si se dice esta variable en la población se puede representar con la curva normal, y que $q_{3/4}$,es decir el tercer cuartil (q_3) es igual a 450 gramos, significa que el 75% de los valores poblacionales son cuando más igual a 450 gramos, y sólo un 25% toma valores por encima. Nótese la equivalencia entre el $q_{3/4}$ y el percentil 75, p_{75} .

Resumen para interpretar los cuantiles

CUARTILES	Primer cuartil, $q_{1/4}$ O bien q_1	deja a su izquierda el 25% de la distribución y el 75% a su derecha
	Segundo cuartil, $q_{2/4}$ o bien q_2	deja a su izquierda el 50% de la distribución y el 50% a su derecha
	Tercer cuartil, $q_{3/4}$ o bien q_3	deja a su izquierda el 75% de la distribución y el 25% a su derecha.

	Primer percentil, $q_{1/100}$ o bien p_1	deja a su izquierda el 1% de la distribución y el 99% a su derecha
	Segundo percentil,	deja a su izquierda el 2 % de la distribución y el 98% a su

Medidas de dispersión.

Las medidas de posición central, por sí solas sabemos que son insuficientes para describir una variable relacionada con un fenómeno de interés, de modo que tengamos una correcta comprensión del mismo. Para reforzar esta idea, recordemos la situación más simple que se nos puede presentar al estudiar una muestra de variables cuantitativas: el caso de muestras pequeñas. Para ellas, vimos que era "obligatorio" utilizar al menos una medida promedio y una medida de la variabilidad.

Amplitud muestral (recorrido o rango), Δ_m o bién A

Es válido lo visto para muestras pequeñas.

Recorrido intercuartílico Δ_q o bién RI

Para evitar situaciones en que los valores extremos anormales distorsionan la realidad más común, esta medida de dispersión absoluta se define como:

Definición 4.28

El "**rango intercuartílico**" es la diferencia entre el tercer cuartil y el primer cuartil.

$$\Delta_q = RI = q_{3/4} - q_{1/4}$$

Las dos medidas de dispersión descriptas, insistimos, adolecen de un gran defecto: no consideran la totalidad de los valores observados, con lo cual es fácil que distribuciones sustancialmente distintas puedan dar las mismas medidas de dispersión al no acusar éstas cambios en la mayoría de los valores de la variable.

Para evitar estos problemas se recurre a la idea intuitiva de medir alejamientos medios, de los valores de la variable a las distintas medidas de posición central de la distribución, y surgen las medidas de dispersión absolutas que se describen a continuación.

Varianza y desviación típica

En variables continuas, x_i es el valor del punto medio que representa a todos los datos clasificados en la clase i-ésima o i-ésimo intervalo de clase.

Varianza

Cálculo por el procedimiento directo	*Cálculo por el procedimiento abreviado.*
$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1}$; *siendo i = 1,2,...,k*	$s^2 = \frac{\sum_{i=1}^{k} x_i{}^2 . n_i - \frac{\left(\sum_{i=1}^{k} x_i . n_i\right)^2}{n}}{n-1}$; *siendo i = 1,2,….,k*

Desviación típica

La **desviación típica**, se obtiene según ya se ha visto como la raíz cuadrada positiva de la varianza

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

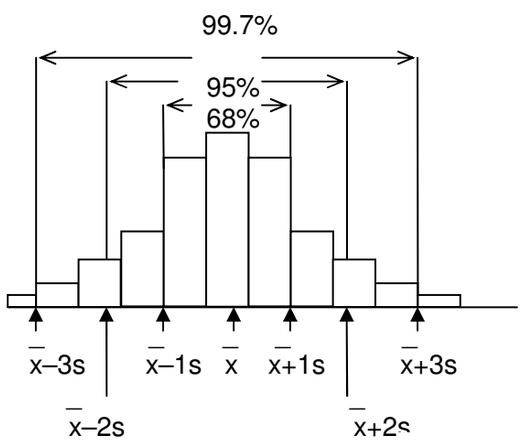
Insistiremos, por la importancia de estas medidas, en su interpretación:

La varianza muestral se puede interpretar como casi un promedio de la suma de cuadrados de desvíos.
La desviación típica, puede ser comprendida examinando dos enunciados:
* la Regla empírica: aplicable a distribuciones de tipo campanular
* el teorema de Chebyshev: aplicable a cualquier distribución

- **Regla empírica**

El examen de muchos conjuntos de datos sugiere una regla empírica que se utiliza para la interpretación de la desviación típica o estándar. Esta regla describe exactamente la variabilidad de los datos poblacionales de una distribución con forma de campana o acampanada, que ya mencionamos es conocida como distribución normal y que se discutirá en detalle en otro capítulo más adelante. Pero también proporciona una descripción bastante adecuada de la variación de muchos otros tipos de variables que poseen distribuciones de frecuencia relativa con forma de pico de montaña.

Además, en la práctica, se puede utilizar la denominada Regla Empírica para explicar la propiedad de variabilidad de los datos de una muestra, esto es: que porcentaje de datos observados se encuentra comprendido por los siguientes intervalos: la media ± k veces la desviación típica. Generalmente estamos interesados en k = 1, 2 ó 3, esto es, la media ± l desviación típica, la media ± 2 desviación típica y la media ± 3 desviación típica, respectivamente, ($\bar{x}$± s) , ($\bar{x}$± 2s) y ($\bar{x}$± 3s) . Estos porcentajes en la muestra se aproximan al 68%, 95% y 99%, respectivamente, en tanto que en la población normal estos porcentajes ocurren de manera exacta (Ver tabla 4.3). La aproximación es tanto mejor, cuanto más grande sea la muestra y además provenga de una distribución normal o aproximadamente normal, es decir, cuando no se da un sesgo extremo y se observa ese aglutinamiento más o menos central de datos. La siguiente Figura muestra los intervalos muestrales comprendidos dentro de una, dos y tres desviaciones típicas de la media y los valores porcentuales el área del histograma abarcada.



Se formalizará ahora el enunciado la regla en discusión.

REGLA EMPÍRICA :
Si una variable está distribuida normalmente, entonces hay un 68% de los datos, aproximadamente, dentro de una desviación estándar de la media. Dentro de dos desviaciones estándares hay un 95% más o menos, y dentro de tres desviaciones estándares de la media hay cerca de 99,7% de los datos. Esta regla es aplicable específicamente a una distribución normal (en forma de campana), aunque con frecuencia se aplica como guía a cualquier distribución de montículo.

Ejemplo 4.8: La regla empírica puede utilizarse para determinar si se puede considerar que la distribución de frecuencias de una muestra aproximadamente se distribuye, o no, de manera normal. Supongamos una muestra, referida a rendimientos en kg/parcela, que tiene una media $\bar{x}$ y una desviación típica s, cuyos valores son 82,9 y 24,3, respectivamente. Utilizando la tabla de la distribución de frecuencias, a través de las frecuencias relativas acumuladas, podríamos encontrar que: el intervalo comprendido desde una desviación típica por debajo de la media hasta una desviación estándar por arriba, esto es $[\bar{x} - s, \bar{x} + s] = [(82,9 - 24,3); (82,9 + 24,3)] = [58,6; 107,2$ comprende el 64% de los datos centrados en la media. Además podríamos encontrar que:

$$[\bar{x} - 2s ; \bar{x} + 2s] = [34,2; 131,5] \quad y \quad [\bar{x} - 3s ; \bar{x} + 3s] = [100 ; 155,8]$$

incluyen el 98% del total de datos y el 100%, respectivamente, de los datos muestrales. Esta información nos lleva a decir que resulta bastante probable que la variable tenga una distribución normal, lo que posteriormente puede ser comprobado a través de la Estadística Inferencial.

- **Teorema de Chebyshev**

La idea asociada al teorema de Chebysheff, para la distribución de datos en una población cualquiera es la siguiente: construir un intervalo fijando una distancia de k a ambos lados de la media μ, con la condición de que k sea por lo menos igual a 1. Entonces, al calcular la fracción $1 - (1/k^2)$, el teorema de Chebysheff afirma que por lo menos esta fracción, del número total de n mediciones, caerá dentro del intervalo determinado.

Tomemos algunos valores numéricos para k. Cuando k=1, el teorema afirma que por lo menos $1-1/(1)^2=0$ de las mediciones caen dentro del intervalo de $\mu-\sigma$ a $\mu+\sigma$, un resultado poco informativo y sin uso práctico, por eso, el teorema resulta útil si k>1. Cuando k=2, resulta que al menos $1-1/(2)^2 = ¾$ de las mediciones caerán en el intervalo $[(\mu -2\sigma) ; (\mu +2\sigma)]$, y cuando k=3, al menos 8/9 de las mediciones estarán en el intervalo de $[(\mu -3\sigma) ; (\mu + 3\sigma)]$, es decir, dentro de tres desviaciones típicas respecto de la media.
Haremos ahora el enunciado formal de la regla en discusión.

TEOREMA DE TCHEBYSCHEFF
La proporción de cualquier distribución situada dentro de k desviaciones estándares de la media es, por lo menos la fracción , $1 - (1/k^2)$, donde k es cualquier número positivo mayor que 1.

Ejemplo 4.9: Ahora consideraremos un ejemplo donde se aplica la media y la desviación típica muestrales, para formar una imagen mental de la distribución de frecuencias para la variable, sin presuponer nada acerca de la población (normal o no). La media y la variación de una muestra con n=25 mediciones, son datos son $\bar{x}=75$ y $s^2=100$. Por lo tanto, la desviación típica es $s=\sqrt{100}=10$. Para una distribución que se centra aproximadamente en $\bar{x}=75$, el teorema de Chebysheff nos permite afirmar lo siguiente:
Al menos ¾ de las 25 mediciones caen en el intervalo $(\bar{x} \pm 2s) = [75 \pm 2(10)]$, es decir, el intervalo de valores x_i que va de 55 a 95.
Al menos 8/9 de las 25 mediciones caen en el intervalo $(\bar{x} \pm 3s) = [75 \pm 3(10)]$, es decir, de 45 a 105.
Finalmente haremos un análisis comparativo, analítico y gráfico, acerca de lo expuesto.

Tabla 4.9: Forma en la que varían los datos alrededor de la media.

Número de desviaciones en unidades k; ($\bar{x} \pm ks$),	Porcentaje de valores de la variable, contenidas entre la media y k desviaciones típicas, para la población	
	Regla de Chebyshev	Distribución de Gauss
k=1	No es calculable	Exactamente 68,26% (≅68%)
2	Al menos 75,00%	Exactamente 95,44% (≅95%)
3	Al menos 88,89%(≅89%)	Exactamente 99,73% (≅ 100%)
4	Al menos 93,75%(≅94%)	Exactamente 99,99%

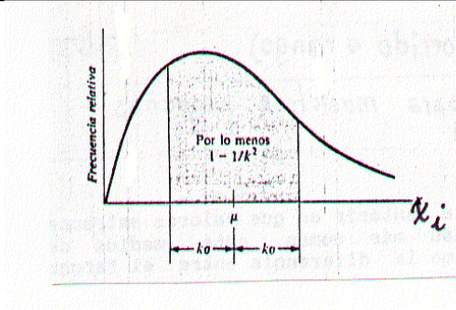


Gráfico 4.6 Ilustración del teorema de Chebyshev

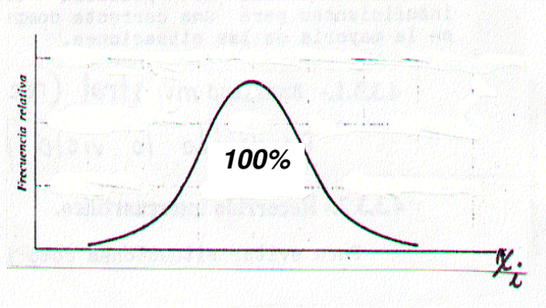


Gráfico 4.7: La distribución campanular

Para concluir, nótese que:
* el teorema de Chebyshev es un hecho que se puede demostrar matemáticamente, y que se aplica a cualquier conjunto de datos (Tabla 4.9 y Gráfico 4.6). Proporciona una cota inferior para la fracción de mediciones que se pueden encontrar en un intervalo ($\bar{x} \pm ks$), donde k es un número mayor que o iguala uno.

* la Regla empírica, por el contrario, es una afirmación arbitraria acerca del comportamiento de los datos. Aunque los porcentajes contenidos en la regla vienen del área bajo la curva normal, los mismos porcentajes son válidos aproximadamente para distribuciones con forma diferente, en tanto tienden a tener forma de pico de montaña (o sea, los datos tienden a acumularse cerca del centro de la distribución).

Coeficiente de variación

Es válido lo visto para muestras pequeñas.

Medidas de asimetría y curtosis.

Hasta ahora, con las medidas de posición hemos situado sobre la escala de valores de la variable las posiciones centrales o más importantes de la distribución y, a través de las medidas de dispersión, hemos medido en promedio el alejamiento o cercanía de los valores de la variable a las medidas de posición central. Sin embargo, aunque a través de estas medidas podemos deducir algo acerca de la "forma" de la distribución de frecuencias, la mayor parte de la información en tal sentido la obtenemos de la observación de las representaciones gráficas de la misma.

Parece, pues, necesario definir una serie de medidas que permitan cuantificar en lo posible la forma de la distribución. Esta cuantificación se realiza en dos sentidos principales:

Propiedad de simetría: simetría o asimetría de la distribución de frecuencias, centrándola en su media, evaluada con las **medidas de asimetría**.
Propiedad de curtosis: la concentración o apuntalamiento más o menos acusada de los valores centrales de la distribución en torno de las medidas de posición central, evaluada con las **medidas de curtosis.**

La simetría y la curtosis, son características propiamente ligada a la forma de la distribución y no a sus valores o unidades de medida. Por ello, cualquier medida que trate de cuantificar exclusivamente algún aspecto de la forma de la distribución debe ser, lógicamente, adimensionales, y en lo posible no influenciables por cualquier transformación de escala o cambio de origen de la escala.

La simetría

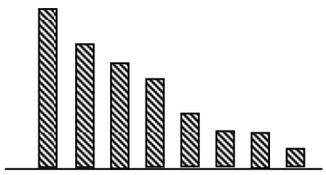
En primer lugar, diremos que vamos a considerar distribuciones unimodales, y que la distribución es simétrica con respecto de algún punto a en el eje de abscisas, si lo es la representación gráfica de sus frecuencias. Es decir, si al trazar una paralela al eje de ordenadas, pasando por el punto a, deja el

mismo número de observaciones a ambos lados, y además, a puntos opuestos y equidistantes de a, siempre les corresponden iguales frecuencias.

Utilizaremos para medirla un coeficiente que se basa en los desvíos a la tercera potencia de los valores de la variable respecto a su media, $(x_i - \mu)$ (o bien para la muestra, $x_i - \bar{x}$), con el siguiente criterio:

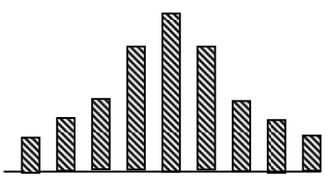
- En una distribución de frecuencias perfectamente simétrica $\bar{x} = \bar{x}_m = \bar{x}_d$, y el índice de asimetría vale cero.
- En una distribución donde $\bar{x} \geq \bar{x}_m$, es decir, la moda es menor que la media, resulta que la distribución se extiende hacia la derecha, tiene exceso hacia valores x_i grandes sesgo positivo, o que "tiene cola hacia la derecha". El coeficiente debería tener signo positivo.
- En una distribución donde $\bar{x} \leq \bar{x}_m$, es decir, la moda es mayor que la media, resulta que la distribución se extiende hacia la izquierda, tiene exceso hacia valores x_i pequeños o sesgo negativo, o simplemente "tiene cola hacia la izquierda". El coeficiente debería tener signo negativo**.**

Así pues los tres casos posibles son:



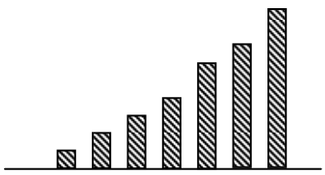
$\overline{x_m} < \overline{x}$

Posición Intervalo modal a la izquierda de $\overline{x_d}$, de $\overline{x}$ y $\overline{x_m}$



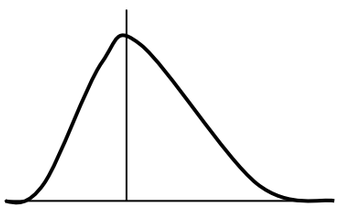
$\overline{x_m} = \overline{x}$

Posición Intervalo modal en coincidencia con $\overline{x_d}$ y $\overline{x}$



$\overline{x} < \overline{x_m}$

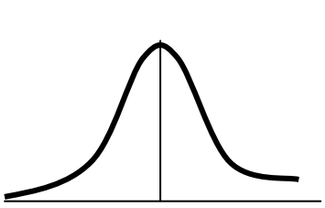
Posición Intervalo modal a la derecha de x_d



$\mu_m < \mu$

Asimetría a la derecha

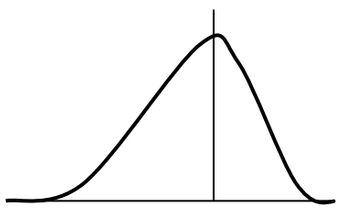
a > 0



$\mu = \mu_d = \mu_m$

Simetría

a = 0



$\mu < \mu_m$

Asimetría a la izquierda

a < 0

En Estadística, la expresión

$$m_r = \frac{\sum (x_i - \mu)^r}{N} \qquad i = 1, 2, \ldots, N$$

corresponde al momento verdadero del **r-ésimo orden**, esto quiere decir desvíos respecto a la media paramétrica.

El momento verdadero de primer orden es igual a cero, $m_1=0$

El momento verdadero de segundo orden resulta ser igual a la varianza poblacional

$$m_2 = \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

El momento verdadero de tercer orden, con desvíos basados en la media y elevados a la tercera potencia $(x_i - \mu)^3$, se relaciona con la propiedad de simetría de un distribución.

El momento verdadero de cuarto orden, análogamente con desvíos $(x_i - \mu)^4$,se relaciona con la propiedad de curtosis.

Para datos muestrales agrupados se tiene la expresión

$$m_r = \frac{\sum_{}^{k} x_i^{\,r} . n_i}{\sum n_i} \quad , \quad i = 1, 2, \ldots, k$$

donde x_i se refiere al desvío entre el i-ésimo punto medio y la media muestral.

Coeficiente de asimetría de Charlier:

$$g_1 = \frac{m_3}{s^3} = \frac{\Sigma\left[\left(x_i - \bar{x}\right)^3 n_i\right] / n}{s^2 . s}$$

donde m_3 son los momentos verdaderos (puntos medios menos la media) de tercer orden, basados en $(x_i - \mu)^3$.

$g_1 < 0$: asimetría negativa; $g_1 = 0$: simetría; $g_1 > 0$: asimetría positiva.

La curtosis

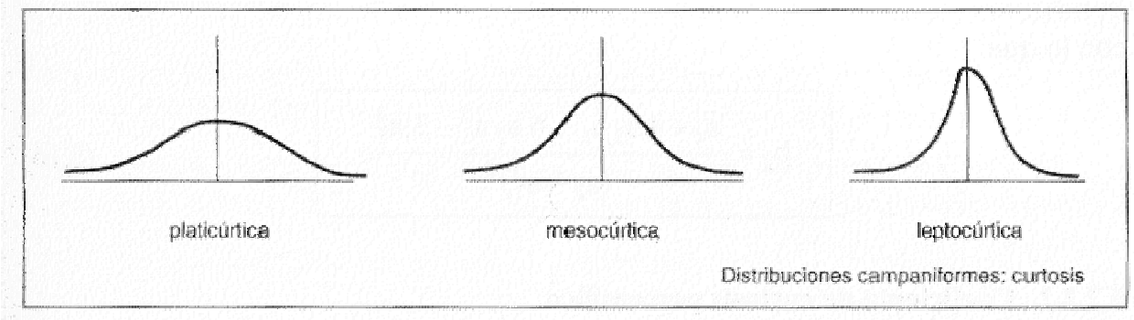
Como ya dijimos, con este coeficiente se trata de medir el grado en que los valores de la variable tienden a agruparse en torno de la media, hay mayor agrupamiento cuanto más elevada o apuntalada sea la distribución. La medida de esta propiedad se hará por referencia a la curtosis que posee la distribución o curva normal, que recibe el nombre de distribución normal, o también campana de Gauss. Esta distribución es simétrica con respecto a su media y verifica que el valor de las potencias cuartas de las desviaciones de las observaciones a la media aritmética, $(x_i - \mu)^4$, vale precisamente tres veces lo mismo que la potencia cuarta de su desviación típica, $3\,\sigma^4$.

Se utiliza para medirla un coeficiente que se basa en la cuarta potencia de los desvíos de los valores de la variable respecto a su media, $x_i - \mu$ (o bien para la muestra, $x_i - \bar{x}$), con el siguiente criterio:

- En una distribución de frecuencias con un grado de concentración similar a la normal, se dirá que la distribución es mesocúrtica y el índice de curtosis debe valer cero.
- En una distribución donde los datos centrales se concentran más que en el caso de la mesocúrtica se dirá que la distribución es leptocúrtica y su índice de curtosis deberá valer más de cero**.**

En una distribución donde los datos centrales se concentran menos que en el caso de la mesocúrtica se dirá que la distribución es **platicúrtica** y su índice de curtosis deberá valer menos de cero.

Gráficamente, las tres situaciones, considerando las poblaciones normales son:



Coeficiente de curtosis:

$$g_2 = \frac{m_4}{s^4} = \frac{\left[\sum_{}^{k}\left(x_i - \bar{x}\right)^4\right] / n}{\left(s^2\right)^2}$$

donde m_4 son los momentos verdaderos (puntos medios menos la media) de cuarto orden basados en $(x_i - \mu)^4$.

$g_2 < 3$: platicúrtica; $g_2 = 3$: mesocúrtica; $g_2 > 3$: leptocúrtica.

El índice se lleva a valor cero para la mesocúrtica, restándole 3 unidades, como:3-3 = 0. De este modo, el índice en una platicúrtica resultará menor a 0 y en una leptocúrtica mayor a 0.

4.9. COMUNICACIÓN Y PRESENTACIÓN DE RESULTADOS

Realizado el análisis estadístico descriptivo (etapa de cálculos) se deberá realizar un informe técnico para comunicar los resultados, en el que se deberán considerar los siguientes aspectos:

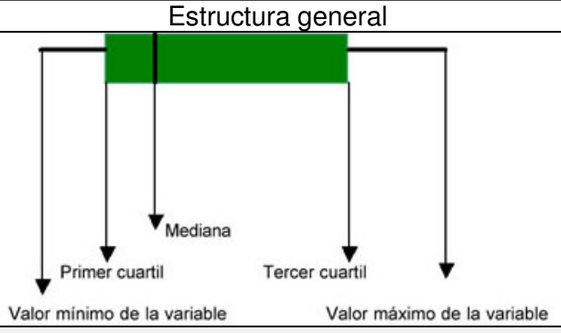
1º) Acerca de los resultados numéricos:

Se deberán redactar conclusiones aplicando la terminología y simbología estadística, y además se deberá proceder a interpretar los resultados en términos del problema. En el caso de variables cuantitativas es muy utilizada la expresión $\bar{x} \pm s$ y, en correspondencia resultados del siguiente tipo: 2,1 *± 0,17,* sin olvidar el acompañamiento de las unidades en que se haya medido la variable.

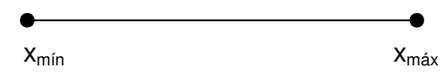
2º) Acerca de los resultados gráficos:

En general se utilizarán gráficos (tablas y representaciones gráficas) de presentación, no corresponde acompañar con tablas auxiliares de cálculo, salvo que se adjunten en un anexo separado.

Además de las representaciones gráficas vistas, conocida la descripción numérica, estamos en condiciones de presentar un nuevo gráfico, el denominado **diagrama de caja** o bien diagrama de caja y bigotes (respectivamente, boxplot y box and whiskers) que es un gráfico muy simple en su forma pero muy informativo en su contenido (describe varias características importantes). El esquema general responde a lo siguiente

Estructura general	Contenido informativo
	Permite visualizar, para un conjunto de datos, información con relación a las cuatro propiedades estadísticas de los datos: a) Posición o tendencia central b) Dispersión general y presencia de datos atípicos. c) Asimetría d) Curtosis

El paso a paso para construir un diagrama de caja es:

1º) Identifique los límites muestrales ($x_{mín}$, $x_{máx}$), posiciónelos en la recta de los reales, y únalos para definir un segmento horizontal (o vertical) con longitud igual a la amplitud muestral,	
2º) Calcule los cuartiles (q_1, $q_2 = \bar{x}_d$, q_3) y posiciónelos en el eje anteriormente trazado. Con los cuartiles 1 y 3 dibuje una caja y particiónela en dos partes trazando una línea en correspondencia al cuartil 2.	
3º) Puede agregar la representación de la media, agregando una cruz	

Ejemplo 4.10. Se trata de construir un diagrama de caja con los datos de una muestra de datos de peso, en kg (n=20)

36	25	37	24	39	20	36	45	31	31
39	24	29	23	41	40	33	24	34	40

1º) Ordenación de los datos

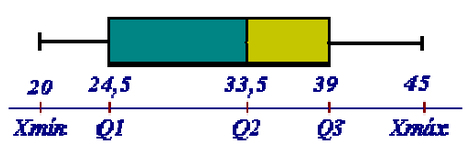
20 23 24 24 24 25 29 31 31 33 34 36 36 37 39 39 40 40 41 45

2º) Identificación de los valores extremos: mín 20 kg y máx 45 kg.

3º) Cálculo de los cuartiles

q_1=(24 + 25) / 2 = 24,5 kg ; $q_2 = \bar{x}_d$ = (33 + 34)/ 2 =33,5 kg; q_3 = (39 + 39) / 2 = 39,0 kg

4º) Dibujar la caja y los bigotes

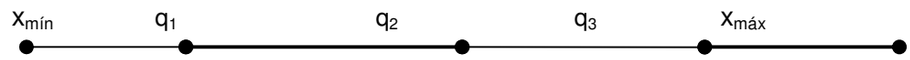


5º) Interpretación gráfica

- El *bigote* izquierdo informa sobre el menor valor de la muestra, y el cuartil 1 (25% de los datos son menores o igual a q_1 y, el 75% son mayores o iguales)

- La caja informa sobre los tres cuartiles: a) el borde izquierdo indica el valor del cuartil inferior y el derecho el valor del cuartil superior, y representa el 50% de los datos posicionados centradamente. La división interna definida por el cuartil mediano, determina dos compartimentos desiguales, cada uno contiene 25% de los datos centrales, pero se observa mayor variabilidad (mayor amplitud) en el primero, y menor variabilidad en el segundo. Nótese también el diferente largo de los bigotes. Puede constatarse en la serie ordenada de datos

20 23 24 24 24 25 29 31 31 33 34 36 36 37 39 39 40 40 41 45



- El *bigote* derecho informa sobre el cuartil 3 (75% de los datos son menores o igual al q_3 y el 25% son mayores o iguales) y el mayor valor de la variable observada en la muestra.

6º) Descripción de las propiedades estadísticas:

Posicionamiento de la distribución:

a1. La mediana tomó el valor 33,5 kg, por tanto un 50% de los datos muestrales correspondieron a pesos menores y un 50% a pesos mayores.

a2. El primero y segundo cuartil, indican que hay un 25% de datos que son inferiores a 24,5 kg (más precisamente, entre 20 y 24,5 kg) y un 25 % que son superiores a 39 kg (más precisamente entre 39 y 45 kg); el 50% restante de los datos presenta valores intermedios a éstos.

a3. Dado que la primera parte de la caja es mayor que la segunda, hay que interpretar que la distribución tiene cola izquierda, con lo cual se induce que el valor de la media es inferior al de la mediana ($\overline{x} < \overline{x_d}$).

Dispersión de la distribución:

b1. Amplitud. $\Delta_m = x_{máx} - x_{mín} = 45 - 20 = 25$ kg; significa que el recorrido total fue de 25 kg, la variación total de la muestra fluctuó entre 20 y 45 kg.

b2. Recorrido intercuartílico. $RI = q_3 - q_1 = 14,5$ kg; es decir, el 50% de los datos muestrales está comprendido entre 24,5 y 39 kg.

b3. Con relación a los valores de la variable comprendidos en el recorrido intercuartílico, se observa que los datos se han distribuido con mayor dispersión a la izquierda de la mediana (la primera parte de la caja es mayor) y menor a su derecha (la segunda parte de la caja se extiende menos). Quiere decir que los pesos variaron más entre 24,3 y 33,5 kg (a la izquierda del valor mediano) que entre 33,5 y 39 kg (a la derecha del valor mediano).

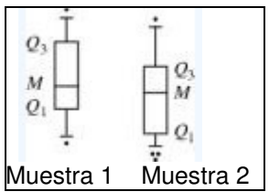
b4. Con relación a los valores comprendidos en los extremos del recorrido total, resulta que el bigote de la izquierda o cola izquierda es más corta que la derecha; por ello el 25% de los pesos inferiores están más concentrados que el 25% de los pesos mayores.

b5. En la muestra no hay datos muy atípicos (desviaciones individuales muy grandes, es decir, pesos excesivamente pequeños o excesivamente grandes). Un valor atípico puede resultar de transponer los dígitos al registrar una medición, de leer incorrectamente la carátula de un instrumento, del mal funcionamiento de una parte del equipo, y de otros problemas. Incluso cuando no hay errores de registro o de observación, un conjunto de datos puede contener una o más mediciones válidas que, por una razón u otra, difieren notablemente de las otras en el conjunto. Estos valores atípicos pueden causar una marcada distorsión en los valores de los estadígrafos, de modo que aislarlos es un paso importante en cualquier análisis preliminar de un conjunto de datos (análisis exploratorio de datos), pero nunca se deberá proceder a simplemente eliminarlos, de hecho los valores atípicos por sí mismos, podrían estar llamando la atención sobre lo siguiente: que contienen información importante no compartida con las otras mediciones del conjunto.

Asimetría de la distribución: este gráfico también proporciona información con respecto a la simetría o asimetría de la distribución general de los datos. Para la interpretación se utilizan los siguientes criterios: a) si la mediana está en el centro de la caja o cerca de él, constituye un indicio de simetría de los datos, b) si la mediana está a la izquierda del centro de la caja o sea se aproxima al primer cuartil, la distribución está sesgada a la derecha (asimetría positiva) y, c) si la mediana está a la derecha del centrode la caja, la distribución está sesgada a la izquierda (asimetría negativa. Asimismo, la longitud relativa de los bigotes se puede emplear como un indicio de su asimetría: el bigote del lado sesgado de la caja tiende a ser más largo que el opuesto. Para el caso de la muestra estudiada, tomando el centrado en la mediana, se observa que las dos partes de la caja tienen diferente tamaño, lo cual indica una falta de simetría. Por ser mayor la primera parte, se interpreta que se trata de una distribución con asimetría negativa (mayor dispersión en cola izquierda).

Curtosis de la distribución: el ancho total de la caja abarca gran parte del recorrido total, por tanto la concentración de los datos no es importante y se trata de una distribución con escaso apuntalamiento, es decir, que es de tipo platicúrtico.

Para finalizar, y comprobar la gran utilidad del diagrama de caja como gráfico resumen de las propiedades estadísticas de los datos en masa, se considerará que se dispone de más de una muestra. Un resultado posible podría ser el siguiente:

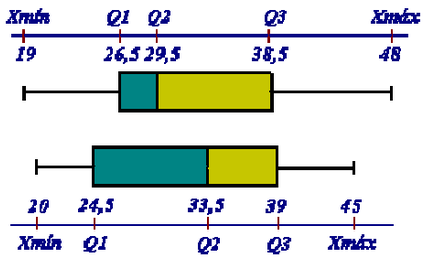


Notar que claramente se muestra que los valores extremos de las muestras son algo diferentes y que la distribución general de los datos también lo es: en la muestra 1 la división de la caja indica mayor variabilidad para los datos por encima de la mediana, mientras que en la muestra 2 ocurre esto con los datos inferiores a la mediana.

Ejemplo 4.11: Supóngase que además de la muestra de pesos analizada precedentemente, se dispone de los datos de una segunda muestra

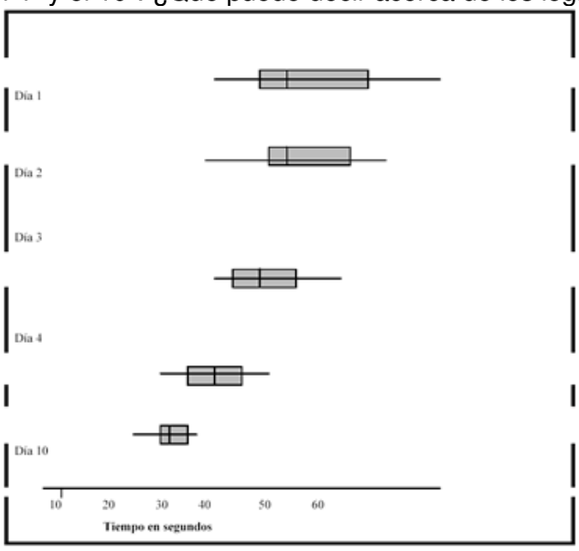
35	38	32	28	30	29	27	19	48	40
39	24	24	34	26	41	29	48	28	22

y al representar los datos del análisis resulta el siguiente diagrama de caja

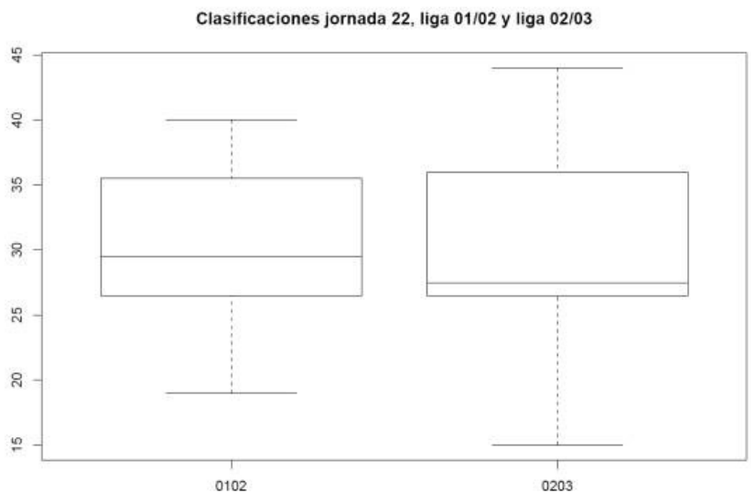


Dejamos al alumno la tarea de realizar un análisis comparativo de las distribuciones correspondientes a los datos de las dos muestras.

Ejemplo 4.12. Intente ahora obtener información acerca de cinco muestras. La variable estudiada es el tiempo que demora un corredor, que se está preparando para una carrera, en recorrer 100m. Su entrenador le ha tomado los tiempos desarrollados en varias corridas realizadas el 1º día de entrenamiento, el 2º, el 3º, el 4º y el 10º. ¿Qué puede decir acerca de los logros del corredor?



Ejemplo 4.13. Seguidamente le proporcionamos dos distribuciones referidas al crecimiento de plantas, en cm, sometidas durante un período de tiempo bajo diferentes condiciones. Realice el análisis comparativo de los resultados.



Para finalizar, cabe explicar el caso de los diagramas de caja que muestran **valores atipicos (outliers).** El cuerpo principal de un diagrama de caja muestra el patrón general de comportamiento que tienen los datos, pero a veces resulta que se tienen algunos datos con un valor "inusual", esto es, datos muy grandes o muy pequeños con relación al patrón general de los datos. Estos datos con valores que se alejan de los restantes pueden deberse a efectos de causas extrañas, como algún error de medición o registro pero también pueden tener otra explicación. Por tanto su eliminación no debe ser precipitada y se justifica recurrir al diagrama de caja para mostrarlos en forma particular. A tal efecto, se requiere agregar otra información al diagrama de caja: la correspondiente a dos tipos de bordes o barreras, internos y externos, que se definen teniendo en cuenta el recorrido intercuartílico (RI), que se calcula como la diferentica entre el cuartil superior y el cuartil inferior del siguiente modo:

Barreras internas	Barreras externas
Barrera interior inferior = Primer cuartil – 1,5 RI Barrera interior superior =Tercer cuartil + 1,5 RI	Barrera exterior inferior = Primer cuartil – 3 RI Barrera exterior superior = Tercer cuartil + 3 RI

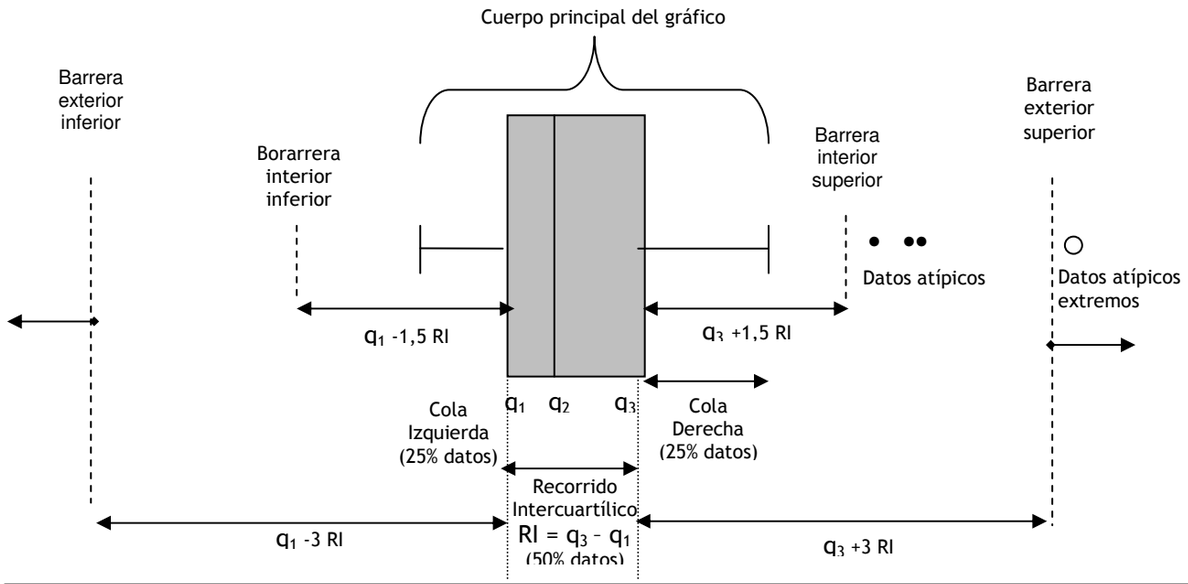


Gráfico 4.8. Diagrama de caja con barreras

Si existen valores de la variable atípicos, según la magnitud de sus desvíos, se los encontrará comprendidos entre las barreras interiores y exteriores.

- Un **valor atípico sospechoso o leve**, se marca en la gráfica con un círculo relleno (•), en cambio un **valor atípico extremo o severo** se suele indicar con un círculo vacío (o) o un asterisco (*).

Los bordes internos y externos se muestran en el gráfico 4.8. con líneas discontinuas, pero usualmente no se dibujan en el diagrama de caja. Cualquier medición que esté entre los bordes interno y externo se llama **valor atípico sospechoso**, y cualquier medición que esté más allá de los bordes externos es u **valor atípico extremo**. Las mediciones que quedan al ubicarse dentro de los bordes, no son raras. El diagrama de caja también marca el rango de las mediciones dentro del borde al localizar los **valores adyacentes**, es decir las mediciones más grande y más pequeña antes de los bordes internos.

Algunas preguntas que conviene formularse para una mejor interpretación y comprensión del comportamiento de la variable observada, son:

- ¿Cuáles son los conceptos del análisis descriptivo (estadígrafos) que pueden analizarse en este tipo de gráfico?
- ¿Qué valores han tomado esos estadígrafos?
- ¿Qué porcentaje de datos representa la caja?
- ¿Qué porcentaje representa cada uno de los bigotes?
- ¿Siempre se encuentra la mediana en el centro de la caja?
- ¿Puede ser un bigote más largo que otro?. ¿Qué estaría indicando, si así fuera?
- ¿Para que sirven las barreras?