1: ESTADÍSTICA DESCRIPTIVA Y ANÁLISIS DE DATOS

La Estadística, en general, trata con información basada en ciertos datos de interés. La palabra "estadística", ha sido referida ya sea a la información misma como a los métodos que tratan con la información. Para evitar confusiones, los estadísticos prefieren llamar a la información: los *datos estadísticos* y a los métodos que tratan con la información: los *métodos estadísticos*.

No toda información es considerada como dato estadístico. Los valores que forman un conjunto de datos estadísticos deben ser tales que se puedan analizar relaciones significativas, es decir, deben ser capaces de ser comparados, analizados e interpretados. Así, la creciente complejidad de las actividades económicas, políticas, científicas, etcétera, ha incrementado el uso de la Estadística para tomar decisiones a todo nivel.

Los *métodos estadísticos* son clasificados en cinco pasos básicos:

- **Recopilación**

  De acuerdo con la localización de la información, los datos estadísticos pueden ser *internos* o *externos*.

  Los datos externos son usualmente obtenidos de dos maneras: de *datos publicados* o de *encuestas o recopilación de primera mano*.

- **Organización**

  El primer paso para organizar un grupo de datos es ordenar y corregir, si es necesario, cada uno de los elementos recopilados.

  El siguiente paso es decidir las clasificaciones adecuadas para incluir todos los elementos.

  El último paso es tabular.

- **Presentación**

  Hay tres modos de presentar un conjunto de datos recopilados: mediante *enunciados* o *textos*, *tablas estadísticas* y *gráficas estadísticas*.

- **Análisis**

  Existen varios métodos de análisis estadístico, mencionamos los más usados:

  - *Análisis estadístico simple*: esta parte proporciona el fundamento básico para el análisis estadístico.
  - *Inducción estadística*: analiza una población o universo basada en un estudio muestral. Otros métodos estadísticos distintos de los inductivos son referidos como Estadística Descriptiva.

- *Análisis de series de tiempo*: analiza los cambios en las actividades de negocios y económicas a lo largo del tiempo.
- *Análisis de relación*: analiza las relaciones entre dos o más conjuntos de datos estadísticos.

- **Interpretación**
  Una conclusión válida puede ser alcanzada después de que los resultados del análisis son interpretados.

Es frecuente que la Estadística se identifique con una tabla o colección de datos, pero no cabe dudas de que la Estadística no debe entenderse como una mera colección de datos, aunque los mismos se presenten de forma ordenada y sistemática.

Como ciencia, la Estadística está formada por el *conjunto de métodos y técnicas que permiten la obtención, organización, síntesis, descripción e interpretación de los datos para tomar de decisiones en condiciones de incertidumbre*.

Para realizar un buen análisis de datos es necesario *organizar* y *sintetizar* para *describir* los datos en estudio.

Veamos cada una de esta etapas:

- **Organización**
  Cuando se compilan datos, deben ser organizados en forma legible.

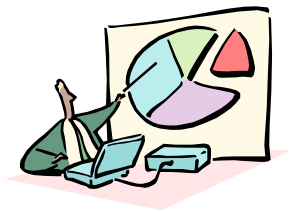
  Pueden ser clasificados en cierta forma sistemática y presentados en un cuadro o tabla.
  Para transmitir su significado más sencilla o más destacadamente, los datos pueden ser presentados en gráficos o diagramas.
  Se calculan luego medidas descriptivas, que permiten 'describir' cuantitativamente los datos y, resumir la información.

- **Síntesis**

  Sintetizar consiste en organizar, comprender, procesar e integrar la información proveniente de múltiples fuentes.
  La síntesis es la reestructuración o reelaboración de la información en formatos nuevos o diferentes para poder cumplir con los requisitos del trabajo.
  La síntesis puede ser tan simple como transmitir un hecho específico o lo bastante compleja, como para contener varias fuentes, varios formatos de presentación o diferentes medios de información y permitir la comunicación efectiva de ideas abstractas.
  En esta etapa es importante enfocarse en comprender la información extraída para presentarla (como producto) en sus propias palabras y en la forma requerida por la tarea.

- **Descripción**

  La descripción de los datos cuantitativos, tales como longitudes, consumos, etcétera, se refiere al cálculo de toda clase de estadísticos (medidas de tendencia central, medidas de dispersión, medidas de posición no centrada, medidas de asimetría, medidas de apuntamiento, entre otras).



  Así mismo, las descripciones se pueden contemplar en modo gráfico, con histogramas, gráficos de tallo y hojas, gráficos de caja y extensiones, diagramas de barras y circulares, con las correspondientes opciones tridimensionales y sus correspondientes propiedades de rotación horizontal y vertical, etcétera.

  La descripción de datos categóricos, tales como zonas geográficas, niveles de aptitud de operarios y alumnos, grados de satisfacción de clientes, etcétera, se realizan mediante efectivos procedimientos de tabulación y tabulación cruzada, que junto con las opciones gráficas, permiten determinar los posibles grados de asociación, entre las categorías analizadas (por ejemplo, la relación entre la afición a la lectura de los padres y el grado de rendimiento escolar de los hijos).

1.1 Presentación de Datos

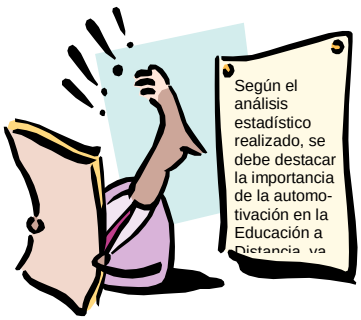
¿Se ha preguntado por qué algunas presentaciones logran captar su atención durante horas, mientras que otras pierden su atractivo en cuestión de minutos?

Veamos algunas pautas que le ayudarán a realizar una presentación efectiva:

- Antes de realizar la presentación, decida qué quiere mostrar y luego haga un esquema de su presentación de principio a fin.
- Tenga en cuenta en qué forma llegará a los interesados la presentación. Por ejemplo: presentación multimedia, presentación impresa, etcétera.
- Esté absolutamente seguro que va a proporcionar información útil para su audiencia, y no sólo lo que usted crea que ella debe recibir de acuerdo con sus conocimientos y experiencia en su área específica.
- Lo que usted quiere es que la audiencia sienta que debe saber, que debe aprender de la información que usted les va a suministrar.
- Al crear su esquema, plantee directamente los puntos principales; luego, respáldelos con investigaciones relevantes, observaciones convincentes y cualquier otro tipo de evidencia que fortalezca el tema de su presentación.
- Use un formato y diseño consistentes. Recuerde que los gráficos deben complementar, no desvirtuar el contenido de la presentación.
- Exprese con claridad sus ideas y conclusiones.

- No base su informe estadístico en una serie de impresiones con salidas de un programa estadístico que carezcan de sentido para el lector.
- Es necesario que realice una interpretación de los resultados obtenidos, e incluso que presente un informe en un lenguaje más cercano a las personas que deben usar los resultados estadísticos sin necesidad de ser expertos.

Existen tres formas para presentar los datos ya organizados y procesados de un estudio estadístico: *texto*, *cuadros* o *tablas* y *gráficas*.



Texto: Esta forma de presentación permite llamar la atención sobre las comparaciones de importancia y destacar ciertas cifras. Sin embargo, sólo puede utilizarse cuando los datos por presentar son pocos.

Cuadros o **Tablas**: Este tipo de presentación permite volcar un gran número de datos en forma resumida, lo que hace fácil y clara su lectura. Además, facilita las comparaciones de los datos.

Al analizar una variable estamos interesados en saber los valores que puede tomar, el número total de datos y cuántas veces aparecen los diferentes valores. La distribución de una variable nos proporciona esta información.

Para presentar variables, tanto cualitativas como cuantitativas, lo podemos hacer mediante una tabla o cuadro, que ofrece una visión numérica sintética y global de dicha variable.

Las tablas o cuadros constan de las siguientes partes:

- El **título**, que debe responder las preguntas: ¿qué?, ¿dónde?, ¿cuándo?
- El **cuerpo**, que consta de: *encabezado de columnas*, *columna matriz*, *columnas secundarias*.
- El **pie de tabla**, que consta de: *fuente de los datos*, alguna *nota* o algún *dato importante.*

Ejemplo:

En un estudio realizado por el Instituto del hierro y el acero de Estados Unidos durante el año 2002, se analizó las cantidades (en miles de toneladas) de importaciones de acero dedicado a la construcción, en distintos países:

Principales fuentes de importaciones de acero en Estados Unidos durante 2002

Países x_i	Frecuencia simple absoluta f_i	Frecuencia simple relativa fr_i	Frecuencia simple relativa porcentual fr_i %
Alemania	460	0,1122	11,22%
Bélgica y Luxemburgo	1247	0,3041	30,41%
Canadá	367	0,0895	8,95%
Francia	299	0,0729	7,29%
Japón	1072	0,2615	26,15%
Reino Unido	250	0,061	6,10%
Otros	405	0,0988	9,88%
	$\sum_i f_i = 4100$	1,0000	100,00 %

Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.
Nota: Para poder operar con los datos de la tabla o referirnos a ella, podemos representar la característica a observar (países) mediante la variable X y a la modalidad i-ésima de dicha variable con la notación x_i.

- **Frecuencia simple absoluta (f_i)**: representa el número de individuos que presentan cada modalidad x_i.
- **Frecuencia simple relativa (fr_i)**: nos permite valorar la representatividad de cada categoría respecto al total de los datos. Se calcula: **f_i / n**.
- **Frecuencia simple relativa porcentual (fr_i%)**: representa en porcentajes las frecuencias simples relativas. Se calcula: **fr_i . 100%**.

Las tablas estadísticas para variables cuantitativas son similares a las anteriores, aunque, en este caso, pueden ser ordenadas con un determinado criterio.

Ejemplo:
Las siguientes son las alturas, en centímetros, de sesenta alumnos universitarios:

150	160	161	160	160	172	162	160	172	151
161	172	160	169	169	176	160	173	184	172
160	170	153	167	167	175	166	173	169	178
170	179	175	174	160	174	149	162	161	168
170	173	156	159	154	156	160	166	170	169
163	168	171	178	179	164	176	163	182	162

Una forma sencilla de organizar los datos se propone en la siguiente tabla:

Estatura de sesenta estudiantes universitarios de Mendoza en 2004

Valores observados x_i	Frecuencia simple absoluta f_i	Frecuencia simple relativa $fr_i = f_i / n$	Frecuencia simple relativa porcentual $fr_i\%$	Frecuencia acumulada absoluta F_i	Frecuencia acumulada relativa $Fr_i = F_i/n$	Frecuencia acumulada relativa porcentual $Fr_i\%$
149	1	0,0167	1,67 %	1	0,0167	1,67%
150	1	0,0167	1,67 %	2	0,0333	3,33%
151	1	0,0167	1,67 %	3	0,0500	5,00%
153	1	0,0167	1,67 %	4	0,0667	6,67%
154	1	0,0167	1,67 %	5	0,0833	8,33%
156	2	0,0333	3,33 %	7	0,1167	11,67%
159	1	0,0167	1,67 %	8	0,1333	13,33%
160	9	0,1500	15,00 %	17	0,2833	28,33%
161	3	0,0500	5,00 %	20	0,3333	33,33%
162	3	0,0500	5,00 %	23	0,3833	38,33%
163	2	0,0333	3,33 %	25	0,4167	41,67%
164	1	0,0167	1,67 %	26	0,4333	43,33%
166	2	0,0333	3,33 %	28	0,4667	46,67%
167	2	0,0333	3,33 %	30	0,5000	50,00%
168	2	0,0333	3,33 %	32	0,5333	53,33%
169	4	0,0667	6,67 %	36	0,6000	60,00%
170	4	0,0667	6,67 %	40	0,6667	66,67%
171	1	0,0167	1,67 %	41	0,6833	68,33%
172	4	0,0667	6,67 %	45	0,7500	75,00%
173	3	0,0500	5,00 %	48	0,8000	80,00%
174	2	0,0333	3,33 %	50	0,8333	83,33%
175	2	0,0333	3,33 %	52	0,8667	86,67%
176	2	0,0333	3,33 %	54	0,9000	90,00%
178	2	0,0333	3,33 %	56	0,9333	93,33%
179	2	0,0333	3,33 %	58	0,9667	96,67%
182	1	0,0167	1,67 %	59	0,9833	98,33%
184	1	0,0167	1,67 %	60	1,0000	100,00%
	n = 60					

Fuente: Datos hipotéticos

- **Variable (x_i)**: para poder operar con los datos de la tabla o referirnos a ella, podemos representar la característica a observar (estatura de los estudiantes universitarios) mediante la variable X y a la modalidad i-ésima de dicha variable con la notación x_i.
- **Frecuencia simple absoluta (f_i)**: representa el número de individuos que presentan cada modalidad x_i.
- **Frecuencia simple relativa (fr_i)**: nos permite valorar la representatividad de cada categoría respecto al total de los datos. Se calcula: $\mathbf{f_i / n}$.
- **Frecuencia simple relativa porcentual ($fr_i\%$)**: representa en porcentajes las frecuencias relativas. Se calcula: $\mathbf{fr_i \cdot 100\%}$.
- **Frecuencia acumulada (F_i)**: representa el número de individuos que presentan una modalidad inferior o igual a x_i . Se obtiene sumando las frecuencias absolutas correspondientes a todos los valores menores o iguales a x_i.

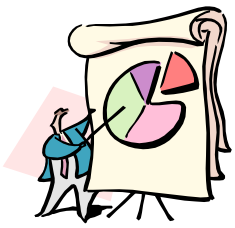
- **Frecuencia acumulada relativa (Fr_i)**: nos permite valorar la representatividad de cada categoría respecto al total de los datos. Se calcula: F_i / **n**.
- **Frecuencia acumulada relativa porcentual (Fr_i%)**: representa en porcentajes las frecuencias acumuladas relativas. Se calcula: Fr_i . **100%**.

Muchas veces, es necesario o resulta más cómodo trabajar con los datos agrupados en intervalos (o clases). La manera de agrupar los datos será estudiada más adelante, por ahora planteamos una posibilidad de agrupación para ver la aplicación en nuestro ejemplo:

Estatura de sesenta estudiantes universitarios de Mendoza en 2004

Intervalos o clases	Punto medio x_i	Frecuencia simple absoluta f_i	Frecuencia simple relativa fr_i	Frecuencia simple relativa porcentual fr_i%	Frecuencia acumulada absoluta F_i	Frecuencia acumulada relativa Fr_i	Frecuencia acumulada relativa porcentual Fr_i%
[149 , 154)	151,5	4	0,0667	6,67%	4	0,0667	6,67%
[154 , 159)	156,5	3	0,0500	5,00%	7	0,1167	11,67%
[159 , 164)	161,5	18	0,3000	30,00%	25	0,4167	41,67%
[164 , 169)	166,5	7	0,1166	11,66%	32	0,5333	53,33%
[169 , 174)	171,5	16	0,2667	26,67%	48	0,8000	80,00%
[174 , 179)	176,5	8	0,1333	13,33%	56	0,9333	93,33%
[179 , 184]	181,5	4	0,0667	6,67%	60	1,0000	100,00%
		n = 60	1,0000	100 %			

Fuente: Datos hipotéticos



Gráficos: La representación gráfica de los datos contenidos en un estudio estadístico tiene como finalidad ofrecer una visión de conjunto del fenómeno sometido a investigación, más rápidamente perceptible que la observación directa de los datos numéricos. De aquí que las representaciones gráficas sean un medio eficaz para el análisis de la información estadística, ya que las magnitudes y las regularidades se aprecian y recuerdan con más facilidad cuando se examinan gráficamente.

La representación gráfica no es más que un medio auxiliar de la investigación estadística, que es fundamentalmente numérica.

Las representaciones gráficas pueden hacerse utilizando un sistema geométrico de representación, en cuyo caso gozan de rigurosidad y precisión, o bien pueden utilizarse símbolos alusivos al tema en estudio (por ejemplo, casas, árboles, figuras humanas, etcétera). Mediante este último sistema de representación

(pictogramas) no se persigue una rigurosa exactitud, sino lograr efectos visuales en quien está leyendo la información.

Existe una gran variedad de gráficos. Su elección depende de las variables en estudio y de las características que se quieren destacar. Para la construcción de gráficos no hay reglas únicas. Siempre se debe tener presente que un gráfico da información más rápida pero menos precisa que una tabla.

1.2 Descripción de un conjunto de datos: Métodos gráficos

A. Datos cualitativos

Aunque una tabla de frecuencias nos proporciona un resumen de datos, en la práctica hay que observar, generalmente, más de un conjunto de datos y compararlos para conseguir una apreciación global y rápida de los mismos. Esto se ve facilitado mediante una adecuada representación gráfica.

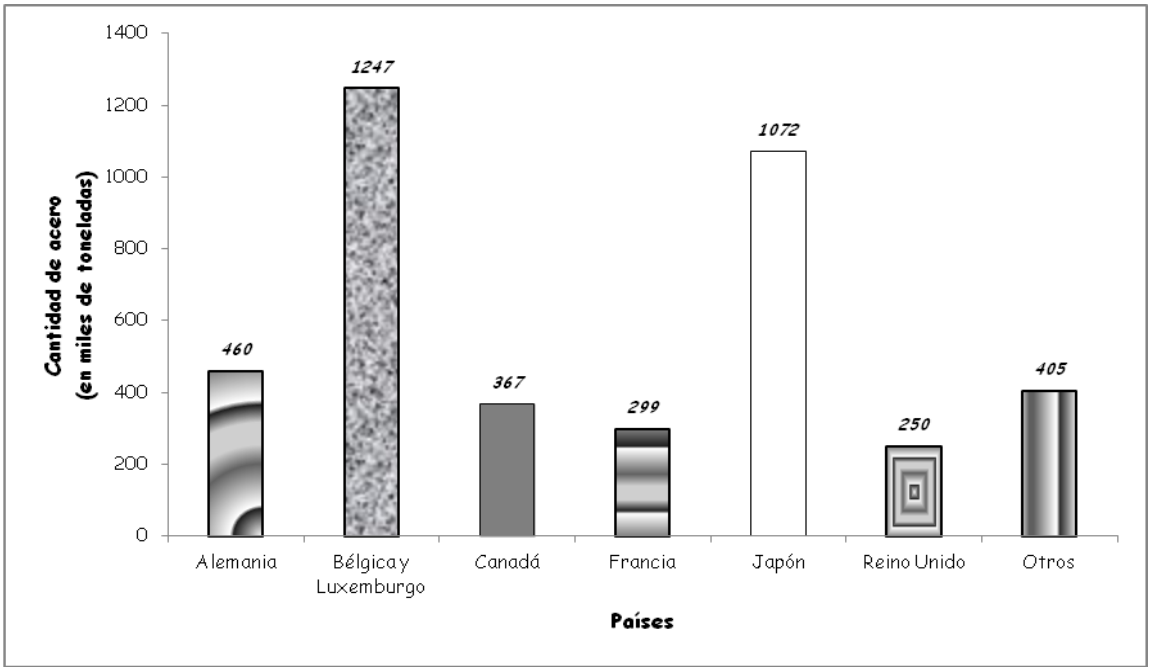
Los gráficos más usuales para variables cualitativas son los ***gráficos de barras***, que pueden ser *verticales* u *horizontales* y los ***gráficos de sectores***.

Ejemplo:

Veremos las distintas representaciones gráficas en un ejemplo anterior:

Gráfico de barras verticales

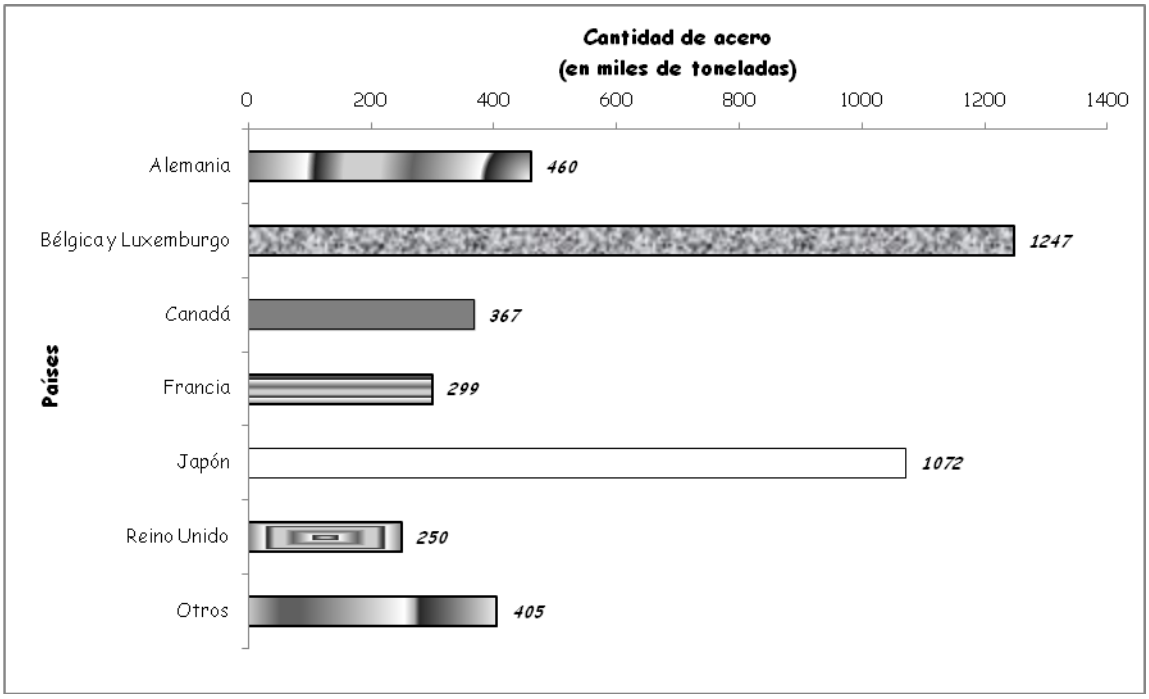
Principales fuentes de importaciones de acero en Estados Unidos durante 2002



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.

Gráfico de barras horizontales

Principales fuentes de importaciones de acero en Estados Unidos durante 2002



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.

Gráfico de Pareto

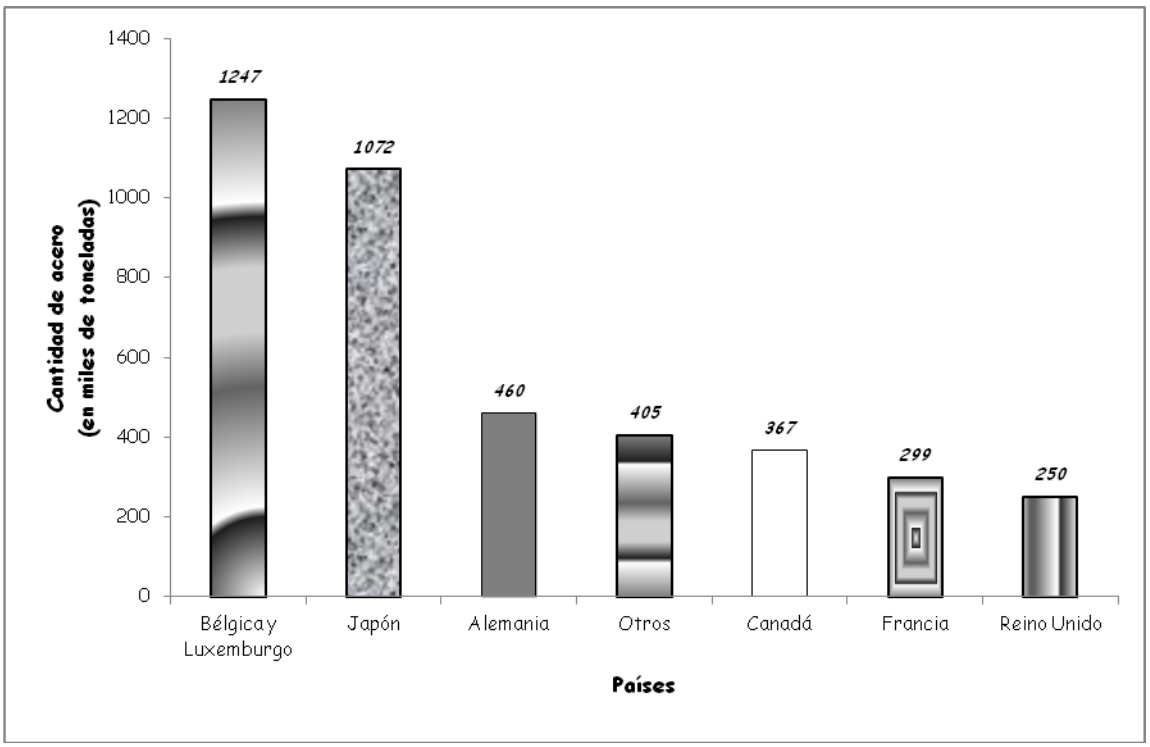
Una variante importante de los diagramas de barras es el llamado diagrama de Pareto. Este diagrama tiene un uso muy amplio, sobre todo, por su valor para realizar comparaciones.

Las categorías están ordenadas de modo tal que en la parte izquierda aparezca la categoría con mayor frecuencia, seguida por la segunda mayor frecuencia y así, sucesivamente. De esta manera, permite asignar un orden de prioridades. Este tipo de diagramas debe su nombre al sociólogo y economista italiano Vilfredo Pareto[1].

El diagrama permite mostrar gráficamente el principio de Pareto (pocos vitales, muchos triviales), es decir, que hay muchos problemas sin importancia frente a unos pocos muy importantes. Mediante la gráfica colocamos los "pocos que son vitales" a la izquierda y los "muchos triviales" a la derecha.

[1] En 1906 hizo la famosa observación de que el 20% de la población poseía el 80% de la propiedad en Italia, posteriormente generalizada por Juran en el principio de Pareto (también conocida como la regla del 80-20).

Principales fuentes de importaciones de acero en Estados Unidos durante 2002



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.

El diagrama facilita el estudio de las fallas en las industrias o empresas comerciales, así como fenómenos sociales o naturales psicosomáticos, como se puede ver en el ejemplo de la gráfica al principio del artículo.

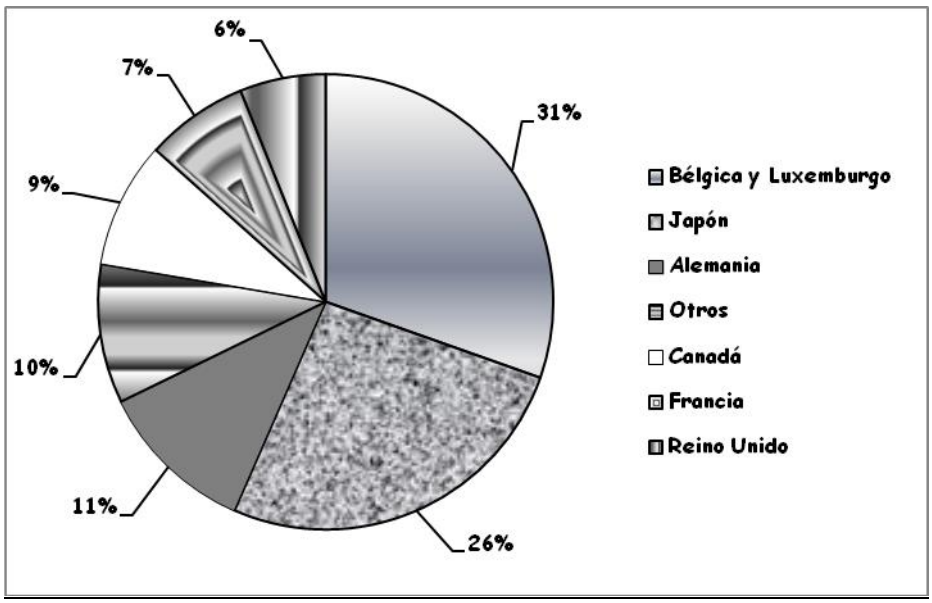
El principal uso que tiene el elaborar este tipo de diagrama es para poder establecer un orden de prioridades en la toma de decisiones dentro de una organización. Evaluar todas las fallas, saber si se pueden resolver o mejor evitarlas.

Gráfico de sectores

Los gráficos de sectores se utilizan para representar variables cualitativas, indicando la proporción en que cada uno de sus valores se presenta.

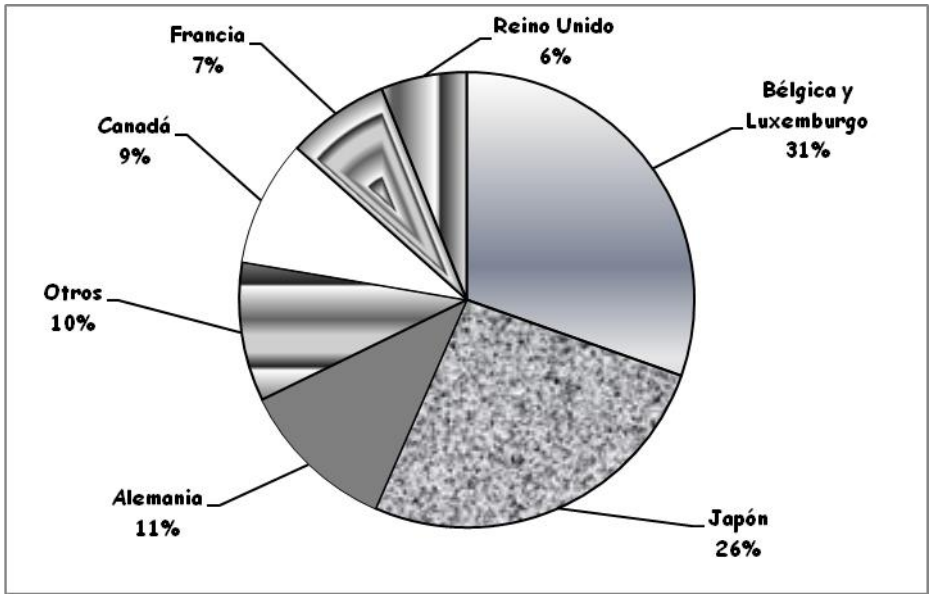
Es usual utilizar distintos colores o tramas para cada sector y colocar las referencias correspondientes (como en el primero de los gráficos de sectores que veremos a continuación). El problema en estos casos es que no siempre se puede apreciar con claridad el color o trama de la referencia. Por esto, suele recomendarse indicar sobre el mismo gráfico la categoría y la frecuencia para ayudar al lector en la comprensión del gráfico (como vemos en el segundo gráfico de sectores).

Principales fuentes de importaciones de acero en Estados Unidos durante 2002



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.

Principales fuentes de importaciones de acero en Estados Unidos durante 2002



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 2002.

B. Datos cuantitativos

Tratamiento de datos individuales

Hay distintas maneras de presentar los datos cuando no han sido agrupados en intervalos. A continuación veremos las más utilizadas:

Gráfico de tronco y hojas

Como hemos visto, es interesante conocer simultáneamente el valor individual de cada una de las observaciones.

El **gráfico de tronco y hojas** (también llamado gráfico de tallo y hojas) fue descrito por Tukey.

Para realizar este gráfico, basta seguir los siguientes pasos:

- Primero se ordenan los datos de menor a mayor.
- Se apartan uno o más dígitos de cada dato, según el número de filas que se desea obtener, en general no más de 15, empezando por la izquierda. Cada valor diferente de estos dígitos apartados, se lista uno debajo del otro, trazando a la derecha de los mismos una línea vertical. Éste es el tronco.
- Para cada dato original se busca la línea en la que aparece su 'tronco'. Los dígitos que nos quedaban los vamos escribiendo en la fila correspondiente de forma ordenada.

Ejemplo:

Se desea analizar cuánto demora un procesador X en guardar un archivo de cierto tamaño. Los tiempos, en segundos, que se recopilaron fueron veinticinco y están dados en la siguiente tabla:

0,8	2,2	0,7	2,6	3,9
2,4	1,2	0,2	0,5	1,2
1,4	3,7	0,4	0,9	1,2
2,1	1,9	0,5	3,8	0,7
1,6	1,4	2,6	0,9	1,5

Entonces, los pasos a seguir son los siguientes:

- Los datos ordenados de menor a mayor son:

0,2	0,4	0,5	0,5	0,7
0,7	0,8	0,9	0,9	1,2
1,2	1,2	1,4	1,4	1,5
1,6	1,9	2,1	2,2	2,4
2,6	2,6	3,7	3,8	3,9

- Observamos que la parte entera de los números son: 0, 1, 2 y 3. Esto nos permite dividir cada número en tronco (la parte entera) y hojas (la parte decimal). Luego, listamos los números que son troncos de arriba abajo y dibujamos una línea vertical.

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

0
1
2
3

- A continuación, para cada dato original, vamos escribiendo, ordenadamente, en el renglón correspondiente al tronco (parte entera), cada una de las hojas (parte decimal). Si alguno se repite, se escribe tantas veces como aparezca. Completando así el gráfico de tronco y hojas.
- Si desea se puede completar el diagrama de tronco y hojas con columnas que indiquen la cantidad de valores que se presentan en cada tronco.

Troncos	Hojas	Frecuencia	Frecuencia relativa
0	2 4 5 5 7 7 8 9 9	9	0,36
1	2 2 2 4 4 5 6 9	8	0,32
2	1 2 4 6 6	5	0,20
3	7 8 9	3	0,12
		n = 25	1,00

Los paquetes estadísticos, en general, presentan este gráfico indicando las frecuencias acumuladas. A continuación veremos un gráfico de tronco y hojas realizado con Statgraphics Plus 5.1 para este conjunto de datos:

```
Stem-and-Leaf Display for Tiempo: unit = 0,1   1|2 represents 1,2

      2      0|24
      9      0|5577899
     (5)     1|22244
     11      1|569
      8      2|124
      5      2|66
      3      3|
      3      3|789
```

El gráfico muestra al conjunto de datos dividido en ocho troncos seguidos de una barra vertical separadora, representados en la segunda columna y seguidos por sus hojas.

Seguramente se preguntará por qué lo hace en ocho troncos si nosotros lo hicimos en cuatro.
No hay un único gráfico de tronco y hojas para un mismo conjunto de datos, es posible realizarlo de distintas maneras, según la necesidad, la claridad e, incluso, la estética que se quiera presentar para la descripción del conjunto de datos.

En este caso, Statgraphics Plus 5.1 propone:

Un tronco para los valores entre 0,0 y 0,4 (0|24); otro para los valores entre 0,5 y 0,9 (0|5577899); otro para los valores entre 1,0 y 1,4 (1|22244); otro para los valores entre 1,5 y 1,9 (1|569); y así sucesivamente hasta los troncos definidos para los valores entre 3,0 y 3,4 (que no tiene hojas porque no se han observado valores en ese intervalo) y entre 3,5 y 3,9 (3|789).

En la primera columna aparecen las frecuencias acumuladas, pero no como estamos acostumbrados, sino que se acumulan desde el menor valor hasta el tronco que contiene al valor que está exactamente en el medio del conjunto de datos (que más adelante estudiaremos y se llama *mediana*) y desde el mayor valor (ubicado en el último renglón) hasta el tronco que contiene a la mediana. La frecuencia correspondiente a este tronco es una frecuencia absoluta simple y se indica entre paréntesis.

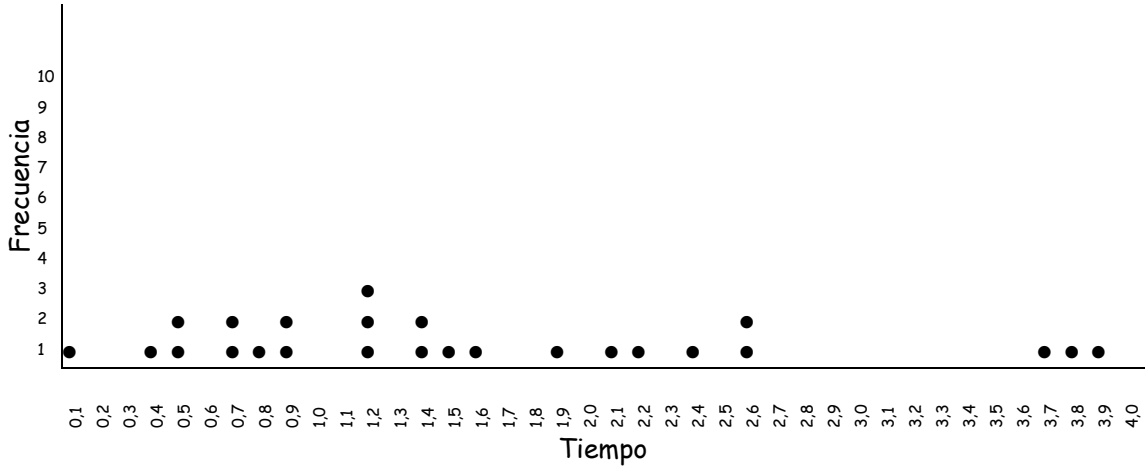
Iremos explicando cómo se han calculado las frecuencias en cada renglón:

2	0|24	Hasta el momento sólo hay **2** valores contados.
9	0|5577899	En este tronco hay 7 valores, más los 2 del tronco anterior, que suman **9**.
(5)	1|22244	Este tronco contiene a la *mediana*, por lo que se registra la frecuencia absoluta simple **5** (hay 5 valores con este tronco) y se la coloca entre paréntesis **(5)**.
11	1|569	La frecuencia presentada es **11** porque es lo acumulado desde el valor más grande (3,9) hasta el primer valor de este tronco (1,5). Es decir, es la cantidad de valores que faltan para llegar al valor más grande.
8	2|124	Son **8** los valores que faltan, desde el primer valor de este tronco (2,1), para llegar a completar los veinticinco datos del conjunto.
5	2|66	Son **5** los valores que faltan, desde el primer valor de este tronco (2,6), para llegar a completar los veinticinco datos del conjunto.
3	3|	Son **3** los valores que faltan, desde el primer valor de este tronco (no hay valores), para llegar a completar los veinticinco datos del conjunto.
3	3|789	Son **3** los valores que faltan, desde el primer valor de este tronco (3,7), para llegar a completar los veinticinco datos del conjunto.

Gráfico de puntos

Para representar gráficamente variables de tipo cuantitativo, si el conjunto de datos es pequeño, usaremos los ***gráficos de puntos*** o ***puntigramas***, que nos permiten distinguir claramente la variable y su frecuencia.

Tiempo de guardado de determinados archivos por un procesador X



Fuente: Datos hipotéticos

Tratamiento de datos agrupados

Tanto las variables discretas como las continuas, con un número grande de valores, se suelen agrupar en intervalos al elaborar las tablas de frecuencias.

Tabla de distribución de frecuencias

Para resumir la información y adquirir una visión global y sintética de la variable en estudio, agruparemos los datos en *intervalos* o *clases*. No obstante, esta operación implica una pérdida de información que será preciso tener en cuenta en la interpretación de las tablas, gráficos y estadísticos de datos agrupados.

La primera decisión que hay que tomar para agrupar una variable es el número de intervalos en que se debe dividir. No existe una regla fija, y en última instancia será un compromiso entre la pérdida de la información que supone el agrupamiento y la visión global y sintética que se persigue. Esta 'flexibilidad' para la selección de la cantidad de intervalos puede provocar dudas o confusiones, es por eso que Sturges da una fórmula para quien no quiera o no pueda decidir la cantidad de clases a utilizar.

Para proceder a la construcción de una distribución de frecuencias con datos agrupados es preciso tener en cuenta las siguientes nociones:

- **Tamaño de muestra (n)**: es la cantidad de elementos en la serie estadística.
- **Máximo ($x_{máx}$)**: se llama máximo de una variable estadística al mayor valor que toma la variable en toda la serie estadística.
- **Mínimo ($x_{mín}$)**: se llama mínimo de una variable estadística al menor valor que toma la variable en toda la serie estadística.
- **Recorrido (R)**: es la diferencia entre el máximo y el mínimo en una serie estadística.

- **Clase**: se llama clase a cada uno de los intervalos en que podemos dividir el recorrido de la variable estadística. Los intervalos pueden o no ser de la misma amplitud.
- **Límite superior de la clase (L_s)**: es el máximo valor de cada intervalo.
- **Límite inferior de la clase (L_i)**: es el mínimo valor de cada intervalo.
- **Marca de clase (x_i)**: es el punto medio de cada clase y es el promedio entre los extremos del intervalo.
- **Cantidad de intervalos (k)**: se obtiene a partir de la fórmula de Sturges, que está dada por: $1+3,3.\log n$. Para tamaños de muestra pequeños también es útil utilizar $\sqrt{n}$ (raíz cuadrada de n), aunque la fórmula de Sturges es válida para todos los casos.
- **Longitud de intervalos (l)**: es la diferencia entre el límite superior y el límite inferior de la clase.

A continuación, aplicaremos un método para dar la distribución de frecuencias de la variable en estudio para datos agrupados.

El método que usaremos, si bien está muy difundido, no es un método único, existen autores y herramientas informáticas que adoptan otros criterios.

Ejemplo:

Analizaremos el ejemplo de las estaturas de los estudiantes universitarios (este conjunto de datos será tomado como ejemplo de aquí en adelante)

150	160	161	160	160	172	162	160	172	151
161	172	160	169	169	176	160	173	184	172
160	170	153	167	167	175	166	173	169	178
170	179	175	174	160	174	149	162	161	168
170	173	156	159	154	156	160	166	170	169
163	168	171	178	179	164	176	163	182	162

El método consta de los siguientes pasos:

- Ordenar los datos de menor a mayor

149	150	151	153	154	156	156	159	160	160
160	160	160	160	160	160	160	161	161	161
162	162	162	163	163	164	166	166	167	167
168	168	169	169	169	169	170	170	170	170
171	172	172	172	172	173	173	173	174	174
175	175	176	176	178	178	179	179	182	184

- Determinar el tamaño de muestra

  $n = 60$
- Reconocer el máximo y el mínimo

  $x_{máx} = 184$ $x_{mín} = 149$
- Calcular el recorrido (también llamado rango o alcance)

  $R = x_{máx} - x_{mín} = 184 - 149 = 35$

- Calcular la cantidad de intervalos

  k = 1 + 3,3.log n = 1 + 3,3 . log 60 ≈ 6,87 ⇒ k = 7 (El valor de k siempre debe ser redondeado a un número entero inferior o superior. Lo usual es hacer el redondeo matemático.)

- Calcular la longitud de cada intervalo

  l = R / k = 35 / 7 = 5 (Si el valor de l resultara ser un número decimal, hay que realizar un redondeo por exceso, con la cantidad de posiciones decimales que se deseen. Por ejemplo, si diera 6,270791, se puede redondear a 6,28 ó 6,3 ó 7, entre otras opciones, pero nunca 6,27 ó 6,2 ó 6.)

- Armar una tabla con los intervalos obtenidos, las marcas de clase y las frecuencias correspondientes

  El primer intervalo está definido desde el valor mínimo (149) hasta 149+5=154. El segundo intervalo está definido desde 154 hasta 154+5=159 y así, hasta el último intervalo que está definido desde 179 hasta 179+5=184.

  Podemos observar que, por ejemplo, el valor 154 estaría en el primero y en el segundo intervalo, por lo que optaremos por definirlos de la siguiente manera: [149;154) - [154;159)-[159;164)-…. De este modo, en caso de existir el valor 154 en nuestro conjunto de datos, será incluido en el segundo intervalo.

  De este modo, con la serie de datos ordenados de mayor a menor, contaremos cuántos valores pertenecen a cada intervalo para completar la tabla.

149	150	151	153	154	156	156	159	160	160
160	160	160	160	160	160	160	161	161	161
162	162	162	163	163	164	166	166	167	167
168	168	169	169	169	169	170	170	170	170
171	172	172	172	172	173	173	173	174	174
175	175	176	176	178	178	179	179	182	184

Intervalos o clases	Punto medio x_i	Frecuencia simple absoluta f_i	Frecuencia simple relativa fr_i	Frecuencia simple relativa porcentual fr_i%	Frecuencia acumulada absoluta F_i	Frecuencia acumulada relativa Fr_i	Frecuencia acumulada relativa porcentual Fr_i%
[149,154)	151,5	4	0,0667	6,67%	4	0,0667	6,67%
[154,159)	156,5	3	0,0500	5,00%	7	0,1167	11,67%
[159,164)	161,5	18	0,3000	30,00%	25	0,4167	41,67%
[164,169)	166,5	7	0,1166	11,66%	32	0,5333	53,33%
[169,174)	171,5	16	0,2667	26,67%	48	0,8000	80,00%
[174,179)	176,5	8	0,1333	13,33%	56	0,9333	93,33%
[179,184]	181,5	4	0,0667	6,67%	60	1,0000	100,00%
		n = 60	1,0000	100 %			

Fuente: Datos hipotéticos

Nota 1: Como el límite superior de cada clase coincide con el límite inferior de la siguiente, adoptamos como criterio que los intervalos se suponen semiabiertos por la derecha, es decir, en cada clase se incluyen los valores de la variable que sean mayores o iguales al límite superior, pero estrictamente menores que el límite superior.

Nota 2: Como excepción al criterio adoptado, en la última clase, el intervalo es cerrado en ambos extremos, si no fuera así, el valor máximo quedaría fuera de los intervalos.

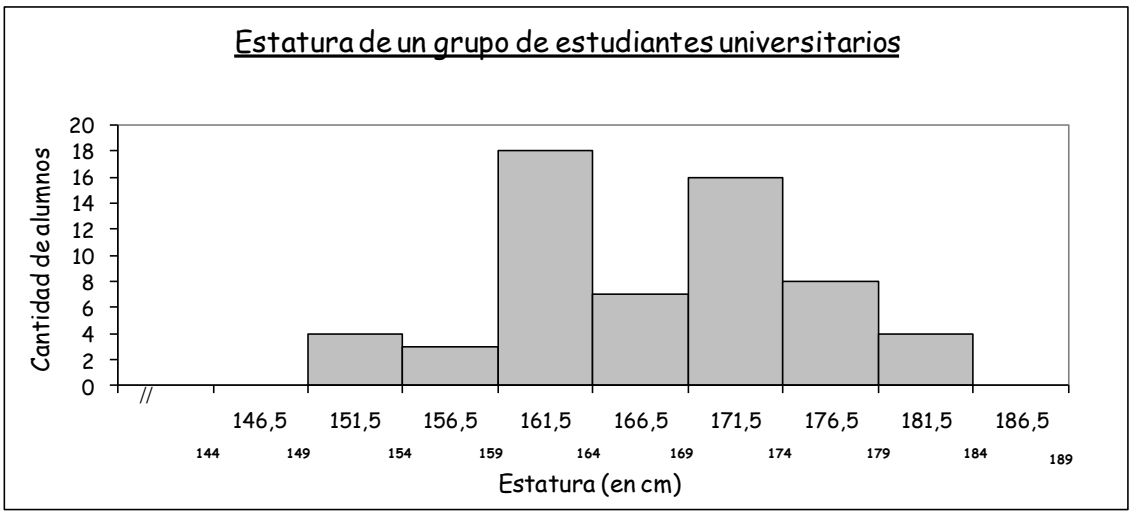
Nota 3: En las frecuencias relativas (fr_i), se debe redondear de tal manera que la suma dé uno.

Histograma y polígono de frecuencias

La información numérica proporcionada por una tabla de frecuencias se puede representar gráficamente de una forma más sintética. En el caso de las variables agrupadas las representaciones que se utilizan frecuentemente son los ***histogramas*** y los ***polígonos de frecuencias***.

Histograma

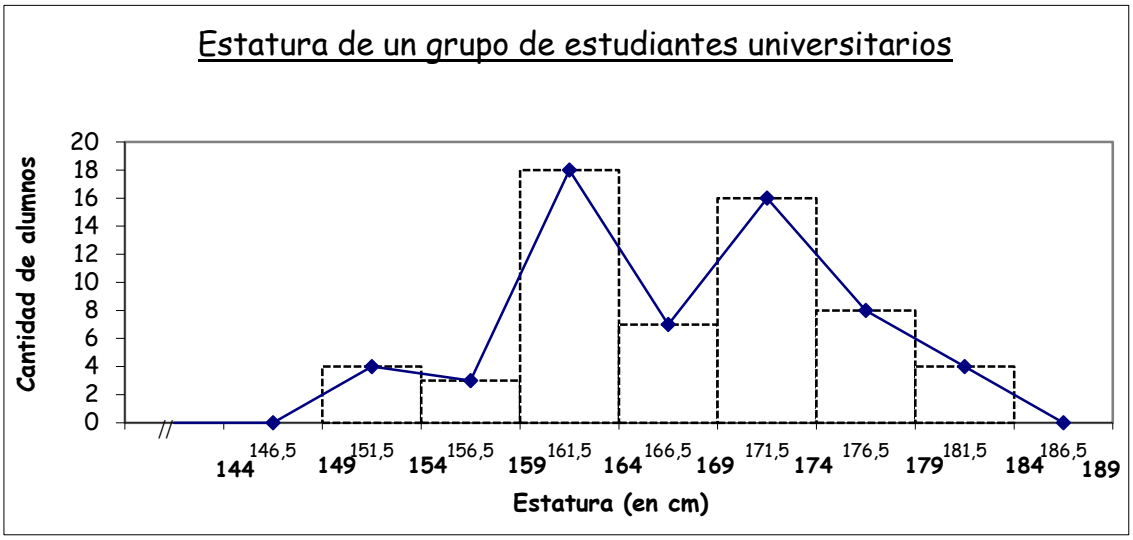
Un histograma se obtiene construyendo sobre unos ejes cartesianos rectángulos cuyas áreas son proporcionales a las frecuencias de cada intervalo.



Fuente: Datos hipotéticos

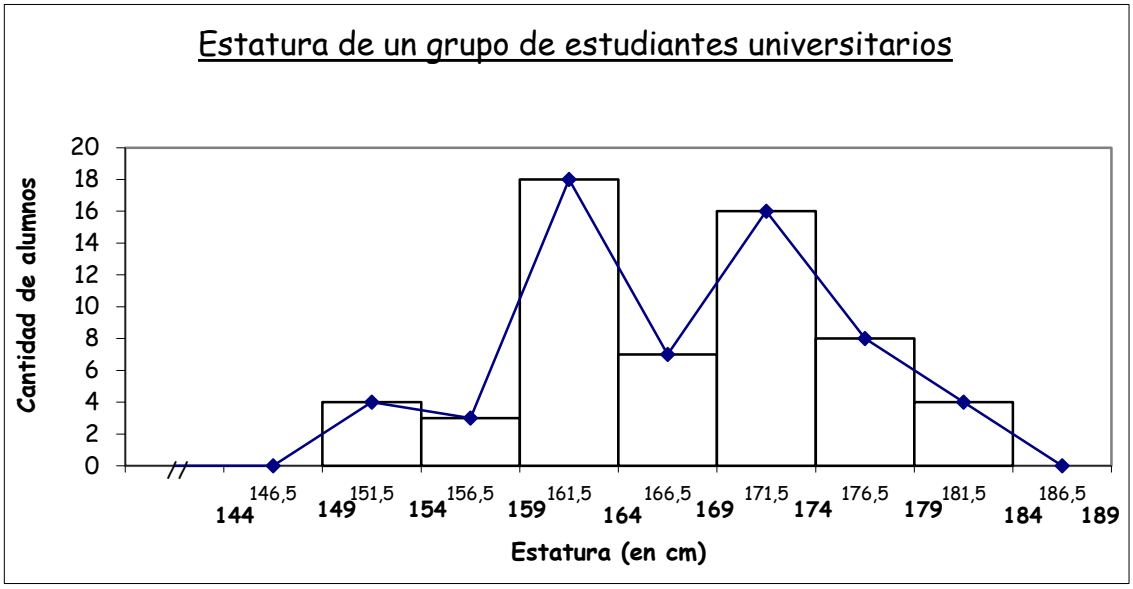
Polígono de frecuencias

Otra forma de representar los datos es el polígono de frecuencias, que es la poligonal que resulta de unir, con segmentos, los puntos medios de las bases superiores de los rectángulos de un histograma de frecuencias.



Fuente: Datos hipotéticos

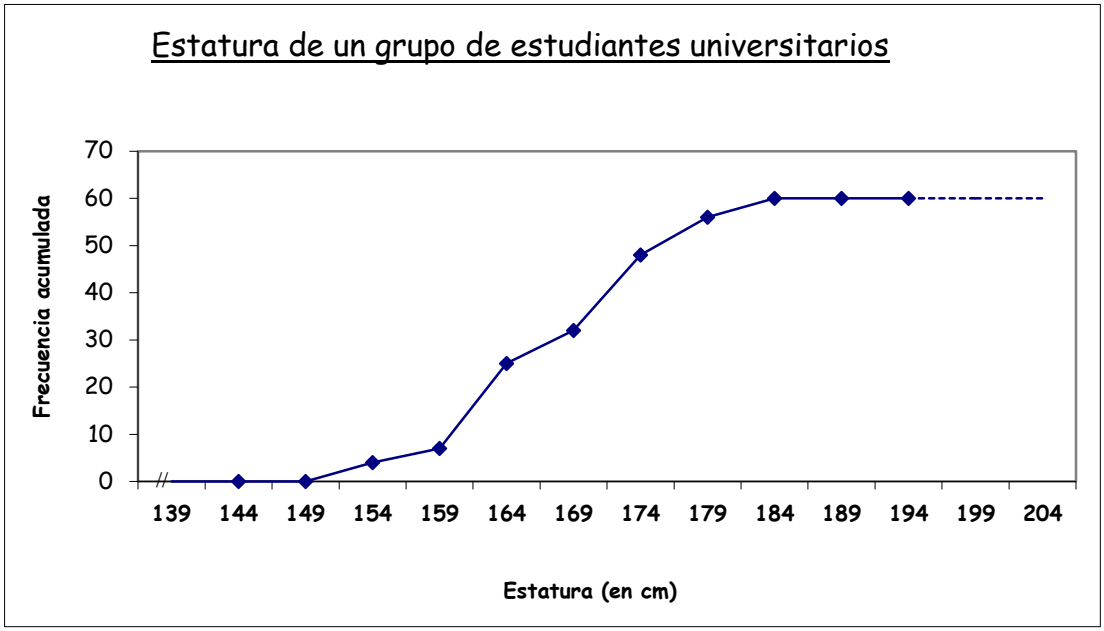
Usualmente se presentan ambos gráficos en el mismo sistema de ejes coordenados.



Fuente: Datos hipotéticos

Ojiva

Llamamos ojiva al polígono de frecuencias acumuladas. Se obtiene uniendo con segmentos los puntos cuyas coordenadas son: la abscisa correspondiente al extremo superior de cada clase y la ordenada correspondiente a la frecuencia acumulada (relativa o absoluta) hasta dicha clase.



Fuente: Datos hipotéticos

C. Patrón de comportamiento

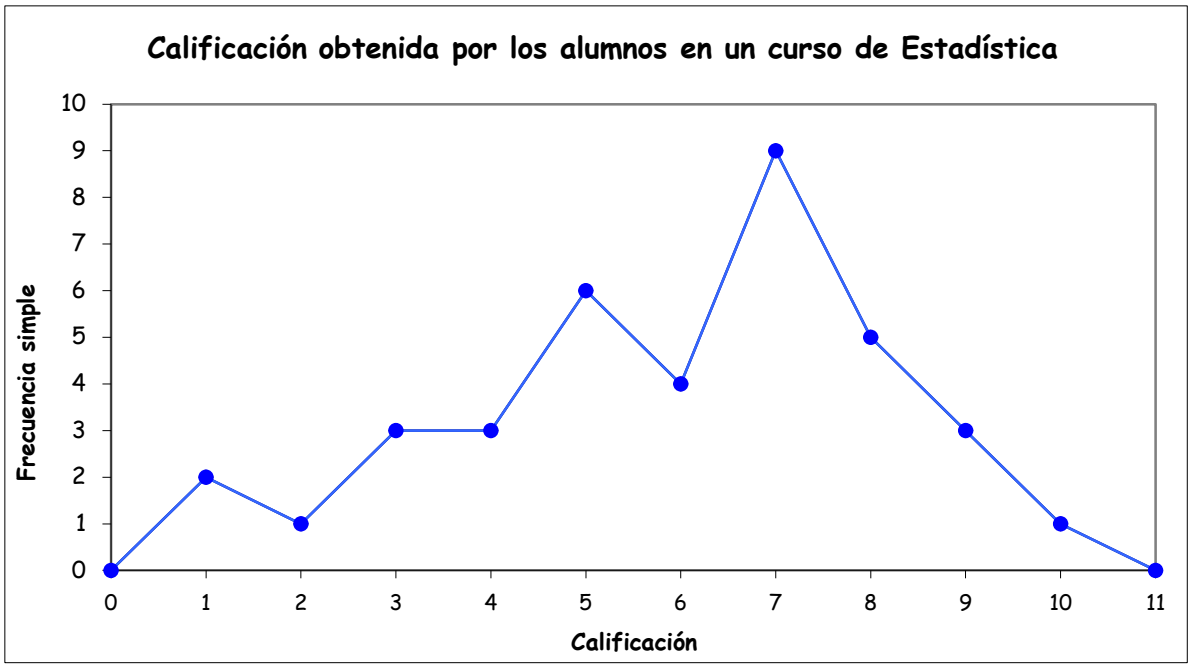
Una tabla de frecuencias, un histograma o un polígono de frecuencias describen una *distribución de frecuencias*, es decir, muestran el patrón de distribución de las frecuencias. En general, las descripciones se refieren a aspectos de la forma del histograma o del polígono de frecuencias.

Un importante aspecto a destacar, relacionado con la forma de una distribución de frecuencias es el hecho de que la figura presente un punto máximo principal. En el ejemplo de las estaturas de los alumnos universitarios, vemos que el intervalo [159, 164) presenta la máxima frecuencia, mostrando un punto máximo en el polígono de frecuencias.

Estos valores máximos se llaman **modos** o **modas** o **valores modales**, en el caso de datos agrupados por clases, a la o las clases que presentan la máxima frecuencia se las llama **clases modales**.

Seguramente se estará preguntando qué ocurre si hay más de un valor o una clase que presente la máxima frecuencia... En este caso decimos que hay dos, tres, ..., k modas. En consecuencia, las distribuciones se llaman **unimodales**, **bimodales**, **trimodales** o **multimodales**, según el número de modas que presenten.

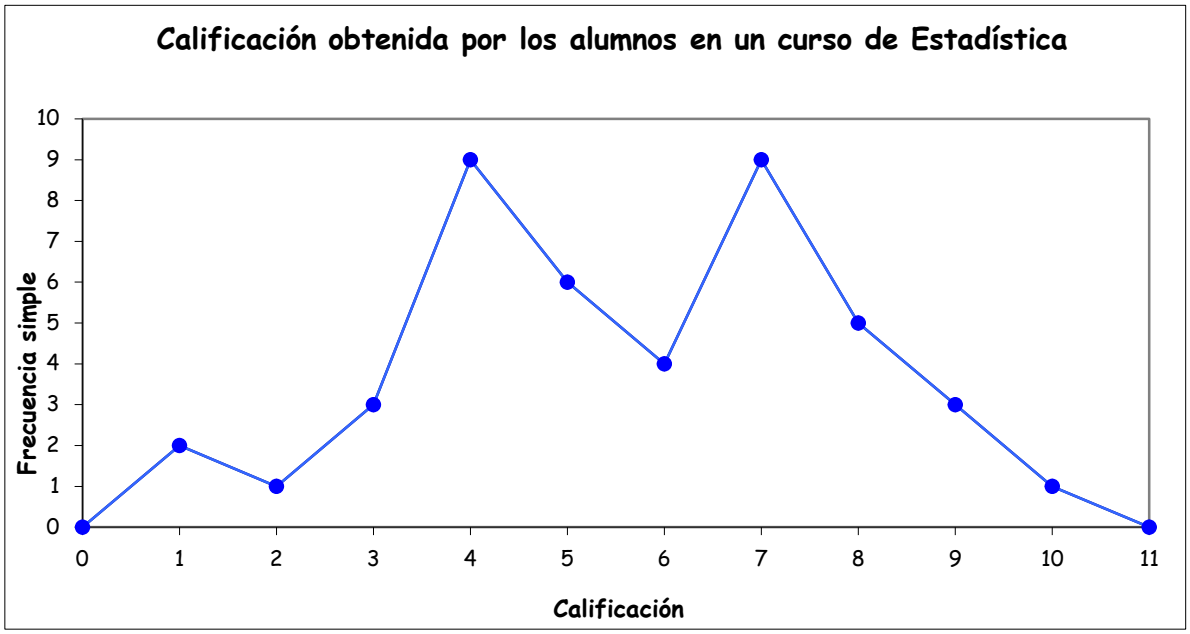
Ejemplo de distribución unimodal:



Fuente: Datos hipotéticos

En este caso la moda es la calificación 7, por ser el valor de la variable que se presenta con mayor frecuencia.

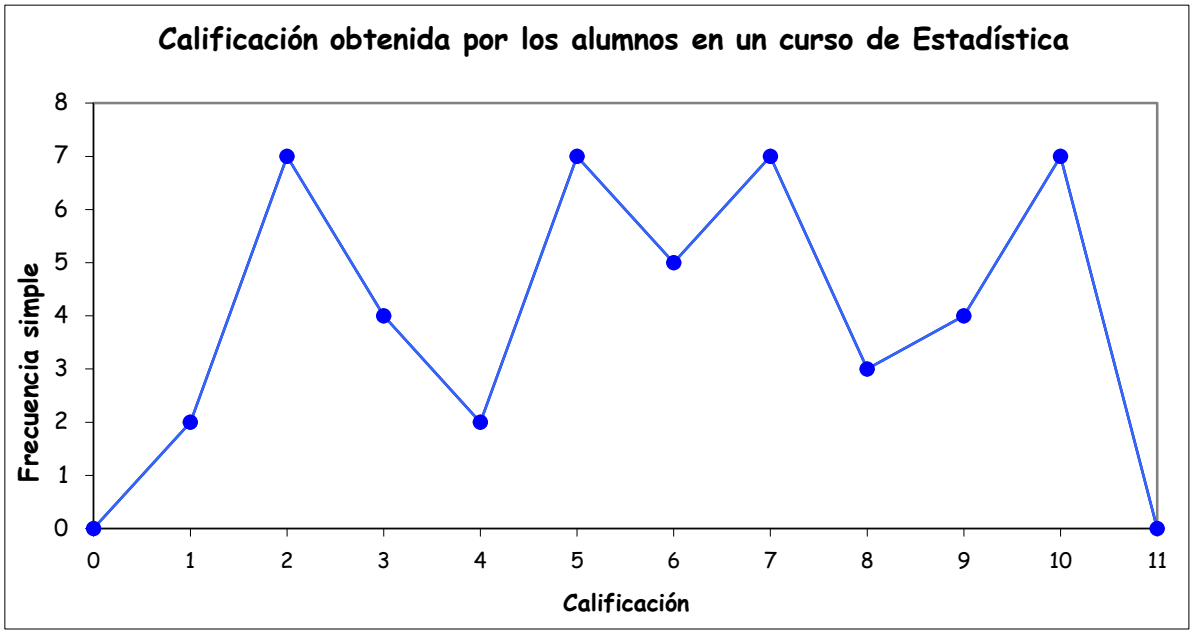
Ejemplo de distribución bimodal:



Fuente: Datos hipotéticos

En este caso se presentan dos modas, las calificaciones 4 y 7, por ser los valores de la variable que se presentan con mayor frecuencia. La distribución es bimodal.

Ejemplo de distribución multimodal:



Fuente: Datos hipotéticos

En este caso la distribución es multimodal y se presentan varias modas, las calificaciones 2, 5, 7 y 10, que son los valores de la variable que se presentan con mayor frecuencia.

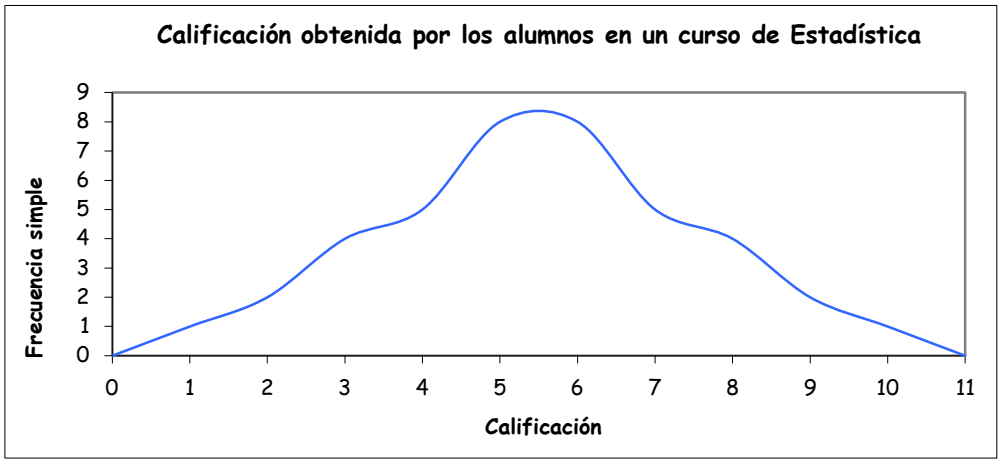
¿Y qué pasa si todos los valores presentan la misma frecuencia...?

En estos casos diremos que...
No hay moda.

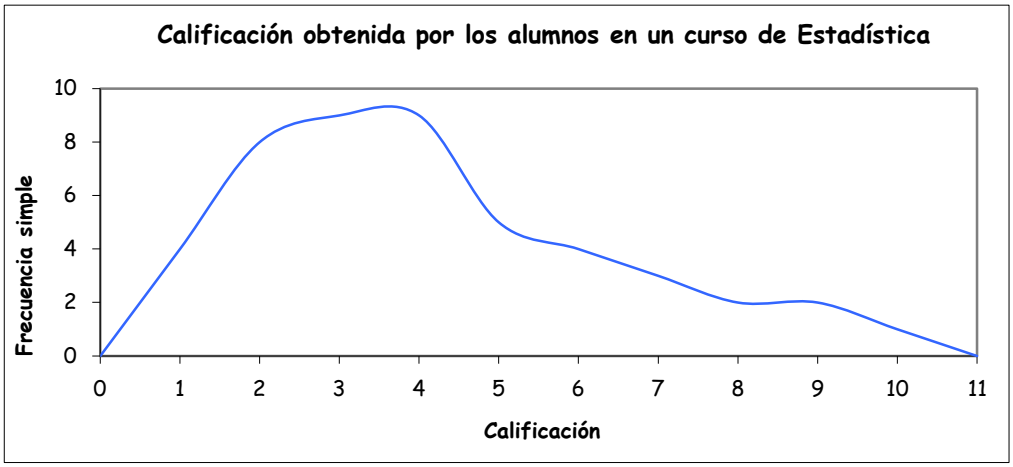
Otro aspecto a destacar, para analizar el patrón de comportamiento de un conjunto de datos, es la **simetría** o **asimetría** de la distribución de frecuencias.

Ejemplos de distribuciones simétricas y asimétricas con distintas asimetrías:

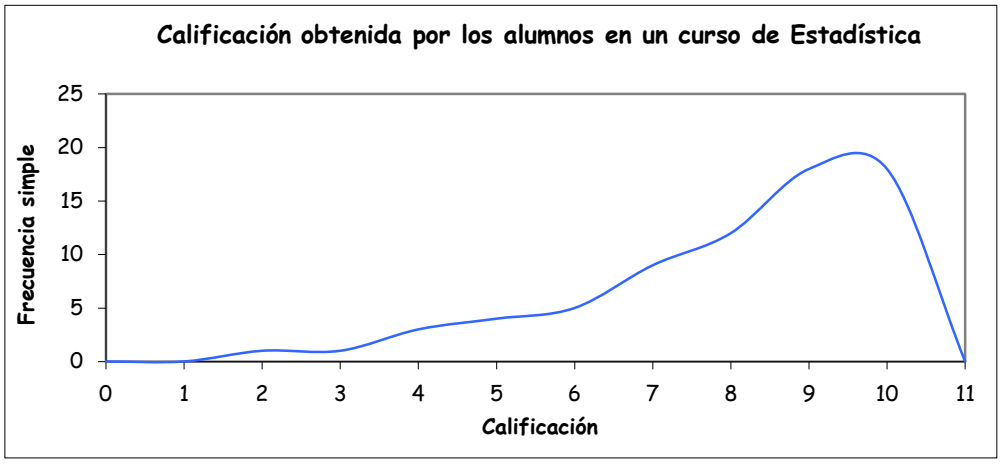
Curso A



Curso B



Curso C



¿Qué podemos decir de cada curso? ¿Cómo se comportan las calificaciones de cada curso? ¿Cuál de los cursos tiene mejor rendimiento?

Claro está que las calificaciones del Curso A presenta una distribución simétrica, mientras que las de los Cursos B y C son asimétricas. La distribución del Curso B se llama **asimétrica a derecha** o **positivamente asimétrica** y la del Curso C, **asimétrica a izquierda** o **negativamente asimétrica.**

Para pensar

- Según el mito popular, ¿qué tipo de distribución tiene la variable: "Cantidad de maniobras que debe hacer una mujer para estacionar correctamente un auto, entre otros dos"?
- A continuación se presentan tablas y gráficos que representan el comportamiento de algunas variables analizadas en el mismo grupo de estudiantes.

Distribución de frecuencias del SEXO de los alumnos

		Frecuencia		Acumulada	
Sexo	Valor	Absoluta	Relativa	Absoluta	Relativa
Hombre	1	17	0.2833	17	0.2833
Mujer	2	43	0.7167	60	1.0000

Gráfico de sectores para la variable:

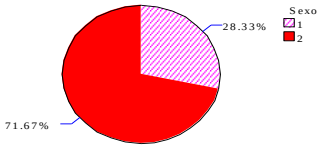


Tabla de frecuencia para la variable DEPORTE

Deporte	Valor	Frecuencia Absoluta	Frecuencia Relativa	Acumulativa Absoluta	Acumulativa Relativa
POCO	1	15	0,2500	15	0,2500
FRECUENTEMENTE	2	32	0,5333	47	0,7833
SISTEMÁTICAMENTE	3	13	0,2167	60	1,0000

POCO: Sólo de vez en cuando
FRECUENTEMENTE: Una vez por semana
SISTEMÁTICAMENTE: Dos o más veces por semana

Gráfico de barras para la práctica de Deporte

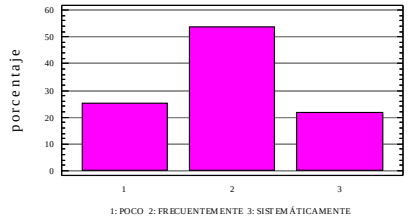


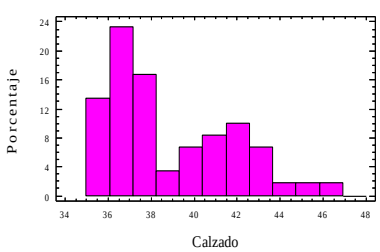
Tabla de frecuencias para PELO y OJOS

	Ojos Claros	Ojos Oscuros	Fila Total
Pelo C (Claro)	17	8	25
	28.33%	13.33%	41.67%
Pelo O (Oscuro)	6	29	35
	10.00%	48.33%	58.33%
Columna Total	23	37	60
	38.33%	61.67%	100.00%

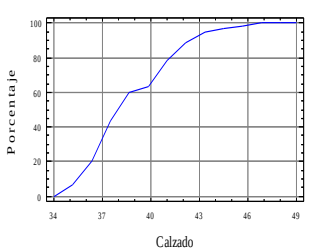
Gráfico de Tallos y Hojas para la variable Número de Calzado de los alumnos

```
Unidad = 0.1   35|0 representa 35.0
      4    35|0000
     12    36|00000000
     26    37|00000000000000
    (10)   38|0000000000
     24    39|00
     22    40|0000
     18    41|00000
     13    42|000000
      7    43|0000
      3    44|0
      2    45|0
      1    46|0
```

Distribución para el Número de Calzado

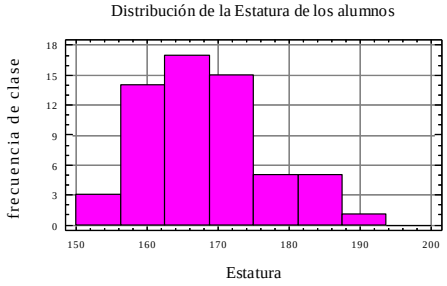


Frecuencias Acumuladas para el Número de Calzado



Distribución de la Estatura de los alumnos (cm)

Clase	Límite de Clase Inferior	Superior	Marca de Clase	Frecuencias Absoluta	Relativa	Acumulativa Absoluta	Relativa
por debajo de		150.0		0	0.0000	0	0.0000
1	150.0	156.25	153.125	3	0.0500	3	0.0500
2	156.25	162.5	159.375	14	0.2333	17	0.2833
3	162.5	168.75	165.625	17	0.2833	34	0.5667
4	168.75	175.0	171.875	15	0.2500	49	0.8167
5	175.0	181.25	178.125	5	0.0833	54	0.9000
6	181.25	187.5	184.375	5	0.0833	59	0.9833
7	187.5	193.75	190.625	1	0.0167	60	1.0000
8	193.75	200.0	196.875	0	0.0000	60	1.0000
sobre	200.0			0	0.0000	60	1.0000



- En base a la observación de las tablas y gráficas, responda:
  - ¿A qué nivel educativo supone que pertenecen estos alumnos?
  - ¿Qué tipos de chistes causarían más efecto, los machistas o los feministas?
  - Respecto a la tabla DEPORTES: ¿Cómo definimos la variable que se refiere a la práctica deportiva? ¿Cómo la codificamos? ¿Cuál es la escala de medición?
  - ¿Cómo se comporta la variable "Número de Calzado"? ¿Cuál es la escala de medición?
  - ¿Es coherente la distribución del número de calzados con el sexo de los estudiantes? ¿Por qué?
  - ¿Qué puede decir respecto al patrón de comportamiento de la variable "Número de Calzado"?
  - ¿Cómo se "comporta" la estatura de los alumnos?

1.3 Descripción de un conjunto de datos: Métodos numéricos

¿Recuerda los conceptos de *población*, *muestra*, *estadísticos* y *parámetros*, que vimos en la Introducción?

Es importante que revise estos conceptos antes de continuar...

Hemos visto que los datos de una muestra pueden ser representados gráficamente, dando una idea global del conjunto de datos analizado.

La representación gráfica de los datos es una primera incursión en el análisis de datos, pero tiene sus limitaciones. Si se desea describir más profundamente el conjunto de datos no siempre es fácil hacerlo a partir de un gráfico e, incluso, no es fácil comparar algunos conjuntos de datos. Por esto, es fundamental *resumir* los datos.

Vimos que podíamos reducir los datos a una forma más compacta, comprensible y comunicable por la distribución de frecuencias.

Las distribuciones de frecuencias no sólo sirven para organizar datos, sino que describen el modelo de distribución de una variable. Realmente, pueden ser consideradas como un conjunto de medidas descriptivas, porque cada número que muestra la frecuencia (o densidad) de observaciones de una clase es una estadística. Pero, a menudo, necesitamos **medidas descriptivas** en forma de números que puedan concentrar mejor la atención en varias propiedades del conjunto de datos que se investiga. Raras veces observamos o medimos poblaciones enteras, por esto, nos dedicaremos a la *descripción de conjuntos de datos*, en términos de *muestras*.

Las características muestrales permiten caracterizar a una muestra con unos pocos valores, llamados *estadísticos*.

Si bien cualquiera función de n observaciones de una muestra es una estadística, hay algunas que son especialmente interesantes. En términos del análisis de datos, nos interesaremos por cuatro propiedades básicas:

- La localización del centro de la distribución, llamadas **medidas de tendencia central**.
- El grado de variación de valores individuales alrededor del punto central o la tendencia de valores individuales a desviarse de las medidas de tendencia central, llamadas **medidas de dispersión**.
- El grado de asimetría, es decir, la falta de simetría de ambos lados del valor modal de una distribución, llamadas **medidas de asimetría**.
- El grado de variación, que representa la proporción de la varianza explicada por la combinación de datos extremos respecto a la media en contraposición con datos poco alejados de la misma, llamadas **medidas de apuntamiento**.

Estas propiedades son significativas especialmente para distribuciones unimodales, pero también se aplican a otros tipos de distribuciones.

A. Medidas de Tendencia Central

Las medidas de tendencia central suelen llamarse *promedios*, y son el 'valor típico' en el sentido de que se emplea a veces para representar todos los valores individuales de un conjunto de datos. Es decir, las medidas de tendencia central dan un valor típico o representativo de un conjunto de datos.

La tendencia central de un conjunto de datos es la disposición de éstos para agruparse ya sea alrededor del centro o de ciertos valores numéricos.

Las más frecuentemente utilizadas son la **media aritmética**, la **mediana** y la **moda**.

Media Aritmética

- La **media aritmética** de las observaciones x_1, x_2, ..., x_n es el *promedio aritmético* de éstas.
- La media aritmética es el valor que tomaría la variable si estuviera uniformemente repartida entre todos los individuos que forman la muestra (corresponde al concepto de centro de gravedad en Física).
- La media aritmética considera todos los datos. Debido a que todas las observaciones se emplean para el cálculo, el valor de la media puede afectarse de manera desproporcionada por la existencia de valores extremos.
- Cuando usemos el término *media*, nos referimos a la media aritmética.

Ventajas de la media aritmética

- Se trata de un concepto familiar para la mayoría de las personas y es intuitivamente claro.
- Cada conjunto de datos numéricos tiene media; siendo ésta una medida que puede calcularse y es única, debido a que cada conjunto de datos posee una y sólo una media.
- La media es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.

Desventajas de la media aritmética

- Aunque la media es confiable en el sentido de que toma en cuenta todos los valores del conjunto de datos, puede verse afectada por valores extremos que no son representativos del conjunto de datos.
- El cálculo se hace tedioso cuando trabajamos con una gran cantidad de valores diferentes.
- Se presentan dudas al calcular la media para clases de extremo abierto, tales como, "mayor que 14" o "menor que 6".
- Sólo se puede calcular para conjuntos de datos numéricos.

Mediana

- La **mediana** es, como su nombre lo indica, el *valor medio* o *valor central* de un conjunto de observaciones.
- Cuando todas las observaciones se ordenan en forma creciente, la mitad de éstas es menor que este valor y la otra mitad es mayor.
- Si el número de observaciones, **n** es **impar**, la mediana es el valor de la observación que se encuentra a la mitad del conjunto ordenado. Si n es impar la mediana es el valor de la observación que ocupa el lugar (n+1)/2.
- Si el número de observaciones, **n** es **par** se considera la mediana como el promedio aritmético, de los valores de las observaciones que ocupan los lugares n/2 y (n+2)/2 del conjunto ordenado.
- Por ejemplo:
  En el conjunto de datos: 5, 3, 8, 2, 7, deberíamos ordenar los datos, o sea, 2, 3, 5, 7, 8, y observar cuál es el valor que está en el medio. Luego, diremos que 5 es la mediana de este conjunto de datos.
  En el conjunto de datos: 5, 7, 8, 1, deberíamos ordenar los datos, o sea, 1, 5, 7, 8, y ver cuál es el valor que está en el medio. Pero no hay un único valor central porque hay un número par de elementos, entonces, diremos que la mediana es el valor promedio entre los dos valores centrales, es decir, entre 5 y 7. Luego, 6 es la mediana de este conjunto de datos.
  En el cálculo de la mediana los valores extremos no afecta su valor.
  En el ejemplo anterior, si en lugar del conjunto de datos 2, 3, 5, 7, 8, tuviéramos el conjunto 2, 3, 5, 7, 8976, la mediana seguiría siendo 5, al igual que en el conjunto -1824, 5, 7, 8, seguiría siendo 6.
- Por lo tanto, si un conjunto contiene valores extremos y un número alto de observaciones, la mediana puede ser una medida de tendencia central mucho más deseable que la media aritmética.

Ventajas de la mediana

- Los valores extremos no afectan a la mediana tan intensamente como a la media.
- La mediana es fácil de entender y se puede calcular a partir de cualquier tipo de datos (excepto datos cualitativos nominales), incluso a partir de datos agrupados con clases de extremo abierto, a menos que la clase mediana sea justamente una de las de extremo abierto, por ejemplo, la clase "mayor que 4".

Desventajas de la mediana

- Ciertos procedimientos estadísticos que utilizan la mediana son más complejos que aquellos que utilizan la media.
- Como debemos ordenar los datos antes de llevar a cabo cualquier cálculo, esto consume mucho tiempo si el conjunto de datos es muy grande.

Modo, Moda o Valor Modal

- La **moda**, **modo** o **valor modal** de un conjunto de observaciones es el valor de las observaciones que ocurre con *mayor frecuencia* en el conjunto.
- El modo es la única medida de tendencia central que puede ser calculada para cualquier tipo de variable (cualitativas o cuantitativas).
- El valor de la moda no se ve afectado por la existencia de valores extremos.
- Puede suceder que en una serie de datos haya más de una moda. En tal caso se denomina bimodal, trimodal o multimodal, según el número de modas que presente.

Ventajas de la moda

- La moda, al igual que la mediana, se puede utilizar como una posición central para datos tanto cualitativos como cuantitativos.
- La moda no se ve mayormente afectada por los valores extremos. Incluso si los valores extremos son muy altos o muy bajos, escogemos el valor más frecuente del conjunto de datos como el valor modal. Podemos utilizar la moda sin importar qué tan grandes o qué tan pequeños sean los valores del conjunto de datos, e independientemente de cuál sea su dispersión.
- Podemos calcular la moda o la clase modal aun cuando una o más clases sean de extremo abierto.

Desventajas de la moda

- A menudo, no existe un valor modal debido a que el conjunto de datos no contiene valores que se presenten más de una vez.
- Cuando los conjuntos de datos contienen muchas modas, resultan difíciles de interpretar y comparar.

Tratamiento de datos agrupados

Ejemplo:

Retomaremos el ejemplo de las estaturas de los alumnos universitarios. En primer lugar, como datos individuales y luego como datos agrupados.

150	160	161	160	160	172	162	160	172	151
161	172	160	169	169	176	160	173	184	172
160	170	153	167	167	175	166	173	169	178
170	179	175	174	160	174	149	162	161	168
170	173	156	159	154	156	160	166	170	169
163	168	171	178	179	164	176	163	182	162

Datos individuales

x_i	f_i	F_i
149	1	1
150	1	2
151	1	3
153	1	4
154	1	5
156	2	7
159	1	8
160	9	17
161	3	20
162	3	23
163	2	25
164	1	26
166	2	28
167	2	30

x_i	f_i	F_i
168	2	32
169	4	36
170	4	40
171	1	41
172	4	45
173	3	48
174	2	50
175	2	52
176	2	54
178	2	56
179	2	58
182	1	59
184	1	60
	n=60	

Media aritmética

Como la media aritmética de las observaciones x_1, x_2, ..., x_n es el *promedio aritmético* de éstas, se denota por:

$$\bar{x} = \frac{\sum_i x_i . f_i}{n}$$

Para datos individuales, los x_i son todos los posibles valores que pueda tomar la variable en estudio y las f_i, las frecuencias absolutas correspondientes.

Trabajando la expresión anterior, $\bar{x} = \frac{\sum_i x_i . f_i}{n} = \sum_i \frac{x_i . f_i}{n} = \sum_i x_i . \frac{f_i}{n}$

Al dividir f_i por n, obtenemos $\mathbf{f_i/n}$, que es la *frecuencia relativa* correspondiente a cada valor x_i. Esta frecuencia relativa es usualmente llamada **peso** de cada valor x_i de la variable estudiada.

Siguiendo la notación del inglés, este *peso* se indica como $\mathbf{w_i}$, por lo que $\mathbf{f_i/n} = \mathbf{w_i}$.

Así, y continuando el trabajo en la expresión de la media aritmética, tenemos:

$$\bar{x} = \frac{\sum_i x_i . f_i}{n} = \sum_i \frac{x_i . f_i}{n} = \sum_i x_i . \frac{f_i}{n} = \sum_i x_i . w_i$$

La media aritmética, definida en función de sus pesos, es llamada **media pesada** o **media ponderada**, quedando expresada como:

$$\bar{x} = \sum_i x_i . w_i$$

En nuestro ejemplo, indicamos la media aritmética como:

$$\bar{x} = \frac{\sum_i x_i \cdot f_i}{n} = \frac{149.1 + \ldots + 166.2 + \ldots + 184.1}{60} = \frac{9993}{60} = \mathbf{166,55\ cm}$$

Interpretación: *La estatura promedio de los estudiantes es de 166,55 cm*

Mediana

Como n es par, para saber la posición del valor de la mediana, buscamos las posiciones n/2 y (n+2)/2, luego, se ven los valores de variable correspondientes y se calcula el promedio entre ellos, obteniendo así el valor de la mediana que deja por encima y por debajo de él, el 50% de las observaciones.
La posición n/2 = 60/2 = 30° corresponde al valor 167 cm
La posición (n+2)/2 = (60+2)/2 = 31° corresponde al valor 168 cm
Luego, la mediana es el valor promedio entre 167 cm y 168 cm, es decir:

$\tilde{x}$ = **167,50 cm**

Interpretación: *El 50% de los estudiantes universitarios observados miden 167,50 cm o menos y el otro 50% miden 167,50 cm o más.*

Modo, moda o valor modal

x_i	f_i	F_i	
149	1	1	
.	.	.	
.	.	.	
.	.	.	
159	1	8	
160	9	17	⇦ Máxima frecuencia absoluta ⇒ Valor modal
161	3	20	
.	.	.	
.	.	.	
.	.	.	
184	1	60	
	n = 20		

El cálculo de la moda para datos individuales es sencillo, basta con buscar el valor de la variable que presente la máxima frecuencia absoluta (f_i).
Luego, la moda es:
Mo= 160 cm
Interpretación: *La estatura de los estudiantes universitarios observados que se presenta con mayor frecuencia es 160 cm.*

Datos agrupados

Intervalos	x_i	f_i	F_i
[149 , 154)	151,5	4	4
[154 , 159)	156,5	3	7
[159 , 164)	161,5	18	25
[164 , 169)	166,5	7	32
[169 , 174)	171,5	16	48
[174 , 179)	176,5	8	56
[179 , 184]	181,5	4	60
		n=60	

Media aritmética

Para datos agrupados basta con extender la definición, considerando a los x_i como los puntos medios de cada intervalo, también llamados marca de clase, y siendo las f_i, las frecuencias absolutas correspondientes a cada clase.

$$\bar{x} = \frac{\sum_i x_i.f_i}{n} = \frac{\mathbf{151,5.4} \quad + \ldots + \mathbf{181,5.4}}{\mathbf{60}} = \frac{10030}{60} = \mathbf{167,17\ cm}$$

Interpretación: *La estatura promedio de los estudiantes es de 167,17 cm*

Mediana

Intervalos	x_i	f_i	F_i	
[149 , 154)	151,5	4	4	
[154 , 159)	156,5	3	7	
[159 , 164)	161,5	18	25	
[164 , 169)	166,5	7	32	⇦ Clase mediana
[169 , 174)	171,5	16	48	
[174 , 179)	176,5	8	56	
[179 , 184]	181,5	4	60	
		n=60		

Para calcular la mediana en datos agrupados seguiremos los siguientes pasos:

- Calcular el orden o posición de la mediana, usando la fórmula (n+1)/2, sin importar si n es par o impar.

  $$°\tilde{x} = \frac{n+1}{2} = 30,5°$$

- Buscar el valor obtenido como orden de la mediana en la columna de frecuencia acumulada (Fi), si no está, tomar el inmediato superior y llamar a la clase correspondiente, *clase mediana*. Diremos que la mediana, $\tilde{x}$, pertenece a este intervalo, pero es necesaria una mayor precisión. Por esto buscaremos el valor de la mediana dentro de la clase mediana.

  $$\tilde{x} \in [164 , 169)$$

- El valor de la mediana se obtiene mediante la fórmula:

$$\tilde{x} = L_{inf\,\tilde{x}} + l \cdot \left(\frac{\frac{n}{2} - F_{ant\,\tilde{x}}}{f_{\tilde{x}}} \right)$$

Siendo:

$L_{inf\,\tilde{x}}$: límite inferior de la clase mediana.

$F_{ant\,\tilde{x}}$: frecuencia acumulada correspondiente a la clase anterior a la clase mediana.

$f_{\tilde{x}}$: frecuencia absoluta correspondiente a la clase mediana.

l : longitud de la clase mediana.

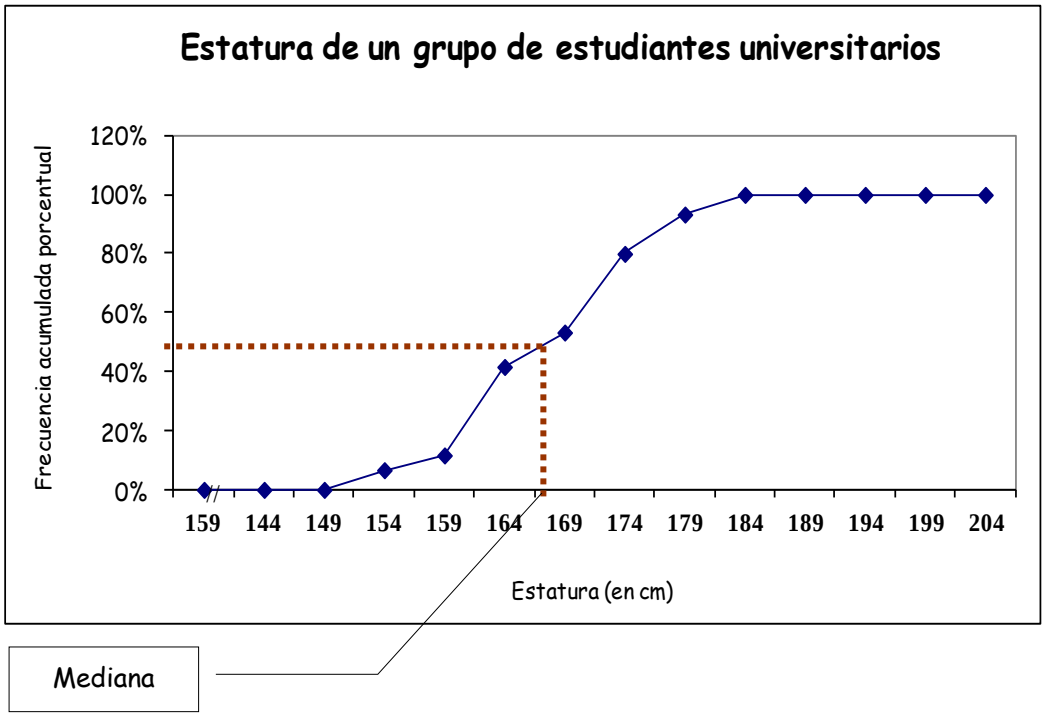
n : tamaño de la muestra.

$$\tilde{x} = L_{inf\,\tilde{x}} + l \cdot \left(\frac{\frac{n}{2} - F_{ant\,\tilde{x}}}{f_{\tilde{x}}} \right) = 164 + 5 \cdot \left(\frac{\frac{60}{2} - 25}{7} \right) = \mathbf{167{,}57\ cm}$$

Interpretación: *El 50% de los estudiantes universitarios observados miden 167,57 cm o menos y el otro 50% miden 167,57 cm o más.*

Nota 1: Otras notaciones para la mediana son: Md, Me y $x_{0,50}$.

Nota 2: La mediana puede calcularse a partir del gráfico de la distribución acumulativa (ojiva), aunque en forma aproximada.
Es conveniente realizar la ojiva colocando en ordenadas la frecuencia acumulada porcentual. Ubicar el 50% y ver a qué valor de abscisa corresponde.



Modo, moda o valor modal

Intervalos	x_i	f_i	F_i	
[149 , 154)	151,5	4	4	
[154 , 159)	156,5	3	7	
[159 , 164)	161,5	18	25	⇦ Clase modal
[164 , 169)	166,5	7	32	
[169 , 174)	171,5	16	48	
[174 , 179)	176,5	8	56	
[179 , 184]	181,5	4	60	
		n=60		

Para calcular la moda en datos agrupados seguimos los siguientes pasos:

- Buscar la máxima frecuencia absoluta y llamar a la clase correspondiente, *clase modal*.
- Diremos que la moda, Mo, pertenece a este intervalo, pero es necesaria una mayor precisión. Por esto buscaremos el valor de la moda dentro de la clase modal.

$$Mo \in [159 , 164)$$

- El valor de la moda se obtiene mediante la fórmula:

$$\mathbf{Mo = x_{Mo} = L_{inf\ Mo} + l \cdot \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)}$$

Siendo:

$L_{inf\ Mo}$: límite inferior de la clase modal

Δ_1 : diferencia entre la frecuencia de la clase modal y la clase premodal (anterior a la modal).

Δ_2: diferencia entre la frecuencia de la clase modal y la clase posmodal (posterior a la modal).

l: longitud de la clase modal

$$\mathbf{Mo} = x_{Mo} = 159 + 5 \cdot \left(\frac{15}{15 + 11} \right) = \mathbf{161,88\ cm}$$

Siendo:

$\Delta_1 = 18 - 3 = 15$ $\Delta_2 = 18 - 7 = 11$

Interpretación: *La estatura de los estudiantes universitarios observados que se presenta con mayor frecuencia es 161,88 cm.*

Para pensar

La siguiente es la distribución de los salarios de los empleados de una pequeña fábrica:

Salario	Cantidad de empleados
$10000	1
$2500	1
$1000	1
$500	2
$200	4

Los empleados realizan una huelga para pedir mejora de sus salarios. Un periodista realiza una nota preguntando cuál es el salario "promedio".
¿Qué medida de tendencia central daría usted si...

a) ... fuera el dueño?
b) ... fuera un representante sindical?
c) ... fuera un investigador científico?

B. Medidas de Dispersión

Las medidas de tendencia central nos indican los valores alrededor de los cuales se distribuyen los datos.

Las medidas de dispersión son estadísticos que nos proporcionan una medida del mayor o menor agrupamiento de los datos respecto a los valores de tendencia central.

Todas ellas son valores mayores o iguales a cero, indicando un valor cero, la ausencia de dispersión.

Para ver sus aplicaciones analizaremos tres muestras de 40 alumnos cada una, a los que se les tomó una evaluación de seis preguntas.

Los x_i indican el número de respuestas correctas y f_i, indica la cantidad de alumnos que lo hicieron.

Grupo 1

x_i	f_i
1	1
2	2
3	17
4	17
5	2
6	1

Grupo 2

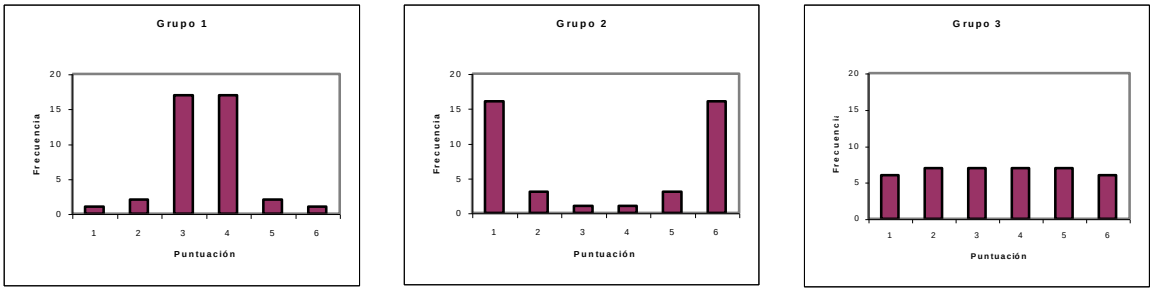
x_i	f_i
1	16
2	3
3	1
4	1
5	3
6	16

Grupo 3

x_i	f_i
1	6
2	7
3	7
4	7
5	7
6	6

Fuente: Datos hipotéticos

Puntuaciones en tres grupos de alumnos



Las tres distribuciones tienen la misma media aritmética y la misma mediana, 3,5 puntos, ¿pero podemos afirmar que hay homogeneidad entre los grupos? Gráficamente vemos que conocer el valor de la media aritmética, de la mediana, e incluso de la moda, no es suficiente para describir cada una de las situaciones.

Para precisar mejor lo que denominamos como 'dispersión' podemos calcular estadísticos que nos den información, sin necesidad de representar los datos.

Rango o Recorrido

- Es la diferencia entre el mayor y menor valor observado de la variable.

$$R = x_{máx} - x_{mín}$$

- El rango indica la variabilidad existente entre las observaciones de un conjunto de datos, sin embargo, debe usarse con precaución, ya que su valor es función únicamente de dos valores extremos pertenecientes al conjunto.
- Debe evitarse el uso del rango como medida de variabilidad, cuando el número de observaciones en un conjunto es grande o cuando éste contenga algunas observaciones cuyo valor sea relativamente grande, respecto al resto.

Varianza

- La **varianza** de las observaciones x_1, x_2, ..., x_n es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto de observaciones.
- El valor de la varianza puede sufrir un cambio muy desproporcionado, aún más que la media, por la existencia de algunos valores extremos en el conjunto de datos.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1}$$

Desviación Estándar

- La raíz cuadrada positiva de la varianza se denomina **desviación estándar** o **desvío típico**.

$$s = + \sqrt{\frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1}}$$

- A menudo se prefiere la desviación estándar con relación a la varianza, porque se expresa en las mismas unidades físicas de las observaciones.
- La desviación estándar nos permite determinar, con un buen grado de precisión, dónde están localizados los valores de una distribución de frecuencias con relación a la media. Podemos hacer esto de acuerdo con un teorema establecido por el matemático ruso P. L. Chebyshev (1821 - 1894).
- El teorema de Chebyshev dice que no importa qué forma tenga la distribución, *al menos* el 75% de los valores caen dentro de $\pm$ 2 desviaciones estándar a partir de la media de la distribución, y al menos 89% de los valores caen dentro de $\pm$ 3 desviaciones estándar a partir de la media.
- Podemos medir aún con más precisión el porcentaje de observaciones que caen dentro de un alcance específico de curvas simétricas con forma de campana. En estos casos, podemos decir que:
  - Aproximadamente 68% de los valores de la población cae dentro de $\pm$ 1 desviación estándar a partir de la media.
  - Aproximadamente 95% de los valores de la población cae dentro de $\pm$ 2 desviación estándar a partir de la media.
  - Aproximadamente 99% de los valores de la población cae dentro de $\pm$ 3 desviación estándar a partir de la media.

Coeficiente de Variación

- Muchas veces nos interesa comparar la variabilidad entre dos o más conjuntos de datos.
- Puede hacerse esto con sus respectivas varianzas o desviaciones estándar cuando las variables se dan en las mismas unidades y sus medias son aproximadamente iguales.
- Cuando no sucede esto, utilizamos una medida relativa de variabilidad llamada coeficiente de variación.
- El **coeficiente de variación** es el cociente entre la desviación estándar y la media aritmética.

$$CV = \frac{s}{\bar{x}}$$

- Esta medida es independiente de las unidades utilizadas.

- El coeficiente de variación es una medida de dispersión relativa, nos indica qué proporción de la media representa la desviación estándar. Por esto, suele expresarse en forma porcentual.
- A partir de la expresión $s = CV \cdot \bar{x}$, podemos interpretar a la desviación estándar en términos de la media aritmética.
- Un inconveniente del coeficiente de variación es que deja de ser útil cuando $\bar{x}$ está próxima a cero.

Tratamiento de datos individuales y agrupados

Ejemplo:

Retomaremos el ejemplo de las estaturas de los alumnos universitarios, en primer lugar, como datos individuales y luego como datos agrupados.

150	160	161	160	160	172	162	160	172	151
161	172	160	169	169	176	160	173	184	172
160	170	153	167	167	175	166	173	169	178
170	179	175	174	160	174	149	162	161	168
170	173	156	159	154	156	160	166	170	169
163	168	171	178	179	164	176	163	182	162

Datos individuales

x_i	f_i	F_i
149	1	1
150	1	2
151	1	3
153	1	4
154	1	5
156	2	7
159	1	8
160	9	17
161	3	20
162	3	23
163	2	25
164	1	26
166	2	28
167	2	30

x_i	f_i	F_i
168	2	32
169	4	36
170	4	40
171	1	41
172	4	45
173	3	48
174	2	50
175	2	52
176	2	54
178	2	56
179	2	58
182	1	59
184	1	60
	n=60	

Rango o Recorrido

R = $x_{máx} - x_{mín}$ = 184 cm – 149 cm = **35 cm**

Interpretación: *La amplitud de la muestra es de 35 cm.*

Varianza

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1} = \frac{(149 - 166{,}55)^2 .1 + \ldots + (166 - 166{,}55)^2 .2 + \ldots + (184 - 166{,}55)^2 .1}{60 - 1} =$$

$$= \mathbf{66{,}18\ cm^2}$$

Interpretación: *El promedio de los cuadrados de las desviaciones de las estaturas respecto a la media aritmética es de 66,18 cm².*

Desviación estándar

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1}} = \mathbf{8{,}14\ cm}$$

Interpretación: *Las estaturas se desvían, en promedio, respecto a la media aritmética, en 8,14 cm.*

Coeficiente de variación

$$CV = \frac{s}{\bar{x}} = \frac{8{,}14\ cm}{166{,}55\ cm} = 0{,}0489 \qquad \mathbf{CV\% = 4{,}89\%}$$

Interpretación: *La desviación estándar representa un 4,89% de la media aritmética.*

Datos agrupados

Intervalos	x_i	f_i	F_i
[149 , 154)	151,5	4	4
[154 , 159)	156,5	3	7
[159 , 164)	161,5	18	25
[164 , 169)	166,5	7	32
[169 , 174)	171,5	16	48
[174 , 179)	176,5	8	56
[179 , 184]	181,5	4	60
		n=60	

Rango o Recorrido

R = $x_{máx}$ - $x_{mín}$ = 184 cm – 149 cm = **35 cm**

Interpretación: *La amplitud de la muestra es de 35 cm.*

Varianza

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1} = \frac{(151{,}5 - 167{,}17)^2 .4 + \ldots + (166{,}5 - 167{,}17)^2 .7 + \ldots + (181{,}5 - 167{,}17)^2 .4}{60 - 1} =$$

$$= \mathbf{63{,}11\ cm^2}$$

Interpretación: *El promedio de los cuadrados de las desviaciones de las estaturas respecto a la media aritmética es de 63,11 cm².*

Desviación estándar

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 . f_i}{n - 1}} = \mathbf{7,94\ cm}$$

Interpretación: *Las estaturas se desvían, en promedio, respecto a la media aritmética, en 7,94 cm.*

Coeficiente de variación

$$CV = \frac{s}{\bar{x}} = \frac{7,94\ cm}{167,17\ cm} = 0,0475 \qquad \mathbf{CV\% = 4,75\%}$$

Interpretación: *La desviación estándar representa un 4,75% de la media aritmética.*

C. Puntuación Z

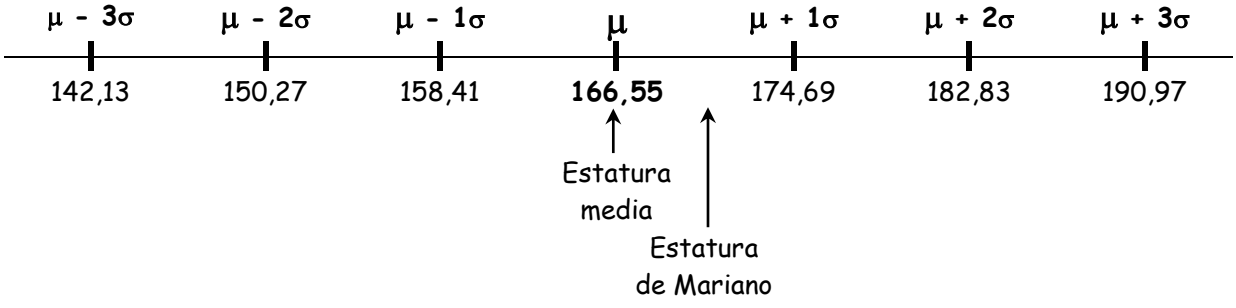
Hasta ahora hemos aprendido a describir una distribución de observaciones en función de la media y la varianza. Ahora aprenderemos cómo describir una observación en particular según el lugar que ocupe dentro del grupo de observaciones en conjunto, es decir, aprenderemos a describir una observación según la misma se encuentre por encima o por debajo del promedio y según a qué distancia por debajo o por encima del mismo esté ubicada.

Ejemplo:

Supongamos que nos informan que Mariano, un estudiante del grupo de alumnos universitarios que venimos analizando, mide 174 cm.

Si desconociéramos las estaturas del grupo sería difícil decir si Mariano es alto o bajo, respecto al grupo de alumnos universitarios. Pero nosotros sabemos que la estatura media es de 166,55 cm y el desvío estándar es de 8,14 cm. Con estos datos, queda claro que Mariano tiene una estatura superior al promedio. También podemos ver que la estatura de Mariano está 7,45 cm por encima de la media.

Supongamos que el conjunto de datos analizado es nuestra población, entonces la media aritmética se debería indicar como μ = **166,55 cm** y la desviación estándar como σ = **8,14 cm**.



¿Qué es una puntuación Z?

Una **puntuación Z** es la transformación de una observación que permite describir mejor el lugar que ocupa esa observación en la distribución.

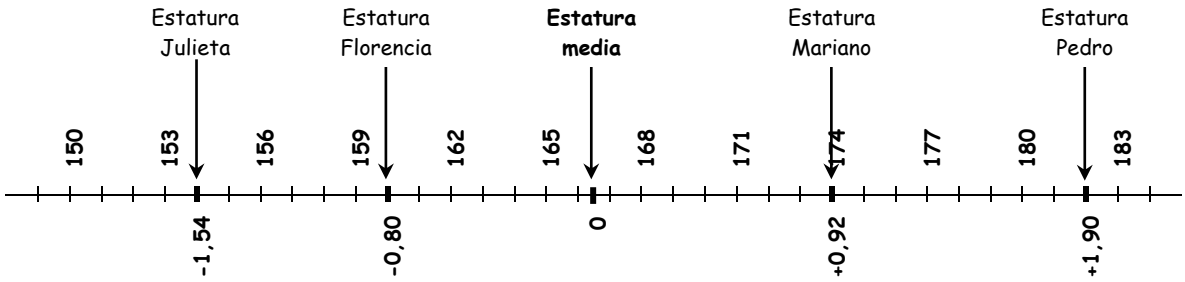
Específicamente, una puntuación Z indica a qué cantidad de desviaciones estándar por encima o por debajo de la media se encuentra dicha observación. Así, el valor Z será positivo si la observación está por encima de la media, será negativo si se encuentra por debajo de la media, y será cero si la observación coincide con la media. La desviación estándar se transforma así en una especie de patrón, una unidad de medida propiamente dicha.

Llamaremos **puntuación bruta** al valor observado, antes de ser convertido en una puntuación Z.

Ejemplo:

Volvamos a nuestro ejemplo de las estaturas de los alumnos universitarios.

- Mariano, que mide **174 cm**, tiene una puntuación Z de **+0,92**, es decir, Mariano está a 0,92 desvíos estándar por encima de la media.
- Florencia, que mide **160 cm**, tiene una puntuación Z de **-0,80**, es decir, Florencia está a 0,80 desvíos estándar por debajo de la media.
- Pedro, que mide **182 cm**, tiene una puntuación Z de **+1,90**, es decir, Pedro está a 1,90 desvíos estándar por encima de la media.
- Julieta, que mide **154 cm**, tiene una puntuación Z de **-1,54**, es decir, Julieta está a 1,54 desvíos estándar por debajo de la media.



¡Sí...! Ya sabemos... Se estará preguntando de dónde sacamos estos valores.

¿Cómo convertir una puntuación bruta en puntuación Z?

Una observación directa se denomina puntuación bruta.

Como hemos visto, una puntuación Z indica la cantidad de desviaciones estándar, por encima o por debajo de la media, a las que se encuentra la puntuación bruta. Para calcular una puntuación Z, se resta la media a la puntuación bruta, obteniendo el desvío. Luego se divide este desvío por la desviación estándar. En símbolos, la fórmula es la siguiente:

$$Z = \frac{X - \mu}{\sigma}$$

Así, si aplicamos la fórmula a las estaturas indicadas en el ejemplo:

- Mariano, que mide **174 cm**, tiene una puntuación Z
  z = (174 cm – 166,55 cm)/8,14 cm = **+0,92**.
- Florencia, que mide **160 cm**, tiene una puntuación Z:
  z = (160 cm – 166,55 cm)/8,14 cm = **-0,80**.
- Pedro, que mide **182 cm**, tiene una puntuación Z:
  z = (182 cm – 166,55 cm)/8,14 cm = **+1,90**.
- Julieta, que mide **154 cm**, tiene una puntuación Z:
  z = (154 cm – 166,55 cm)/8,14 cm = **-1,54**.

¿Cómo convertir una puntuación Z en puntuación bruta?

Dada una puntuación Z, podemos convertirla a la puntuación bruta correspondiente. La obtención de la fórmula es muy sencilla a partir de la dada anteriormente:

Si $Z = (X - \mu)/\sigma$, entonces $\mathbf{X = Z.\ \sigma + \mu}$

Algunas características de las puntuaciones Z

- La puntuación Z de la puntuación bruta correspondiente a la media es 0.
- La puntuación Z de la puntuación bruta correspondiente al valor $\mu - \sigma$ es –1.
- La puntuación Z de la puntuación bruta correspondiente al valor $\mu + \sigma$ es +1.
- Por lo tanto, la puntuación Z tiene media 0 y desviación estándar 1.
- Una gran ventaja de las puntuaciones Z es que, convirtiendo las observaciones (puntuaciones brutas) en puntuaciones Z, podemos comparar variables completamente diferentes entre sí.

D. Medidas de Posición No Centradas

Los estadísticos de orden o medidas de posición no centradas, son aquellos valores numéricos que nos indican su posición en el conjunto de datos ordenados, pues una fracción dada de los datos presenta un valor de la variable menor o igual que el estadístico.

Si una serie de observaciones se colocan en orden creciente, el valor que divide al conjunto de datos en dos partes iguales es la mediana. Por extensión, si preferimos tener una descripción más detallada de la variabilidad de los valores individuales, se puede dividir los datos en otra cantidad de partes iguales. Por ejemplo, en cuatro, en diez o en cien partes iguales, llamando a estas medidas **cuartiles**, **deciles** y **percentiles**, respectivamente.

Cuartiles

- Al dividir los datos en cuatro partes iguales, quedan definidos los **cuartiles**: Q_1, Q_2 y Q_3.
- La fórmula para obtener el lugar del k-ésimo cuartil, siendo n el número de observaciones, es: $^{\circ}Q_k = k.(n+1)/4$ y así, buscando en la lista ordenada de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente. En caso que $^{\circ}Q_k$ no sea un valor entero se calcula por interpolación lineal el valor del cuartil.
- La mediana es el cuartil 2.

Deciles

- Al dividir los datos en diez partes iguales, quedan definidos los **deciles**: D_1, D_2, ..., D_9.
- La fórmula para obtener el lugar del k-ésimo decil, siendo n el número de observaciones, es: $^{\circ}D_k = k.(n+1)/10$ y así, buscando en la lista ordenada de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente. En caso que $^{\circ}D_k$ no sea un valor entero se calcula por interpolación lineal el valor del decil.
- La mediana es el decil 5.

Percentiles

- Al dividir los datos en cien partes iguales, quedan definidos los **percentiles**: P_1, P_2, ..., P_{99}.
- La fórmula para obtener el lugar del k-ésimo percentil, siendo n el número de observaciones, es: $^{\circ}P_k = k.(n+1)/100$ y así, buscando en la lista ordenada de los valores o en la columna de la frecuencia acumulada, se ve el valor de la variable correspondiente. En caso que $^{\circ}P_k$ no sea un valor entero se calcula por interpolación lineal el valor del percentil.
- La mediana es el percentil 50.
- El primer cuartil es el percentil 25.
- El tercer cuartil es el percentil 75.
- El cuarto decil es el percentil 40.
- El ..... decil es el percentil 70.
- El octavo decil es el percentil ..... .

Para muchos problemas es de gran utilidad determinar el recorrido entre dos valores cuantiles que entre dos valores extremos:

- La diferencia entre los percentiles 75 y 25, es decir, entre el tercer y primer cuartil, recibe el nombre de **recorrido intercuartil** y sólo incluye el 50% central de la distribución.
- La diferencia entre los percentiles 90 y 10, es decir, entre el noveno y primer decil, recibe el nombre de **recorrido interdecil** y toma el 80% central de la distribución.

Tratamiento de datos individuales y agrupados

Ejemplo:

Retomaremos el ejemplo de las estaturas de los alumnos universitarios, en primer lugar, como datos individuales y luego como datos agrupados.

150	160	161	160	160	172	162	160	172	151
161	172	160	169	169	176	160	173	184	172
160	170	153	167	167	175	166	173	169	178
170	179	175	174	160	174	149	162	161	168
170	173	156	159	154	156	160	166	170	169
163	168	171	178	179	164	176	163	182	162

Cuartiles, Deciles y Percentiles

El cálculo de los estadísticos de orden, para datos individuales, se ajusta al siguiente método:

Datos individuales

- Calcular el orden o posición del estadístico de orden que se desea obtener.
- Buscar en la serie estadística ordenada en forma creciente, el valor de la variable correspondiente a esta posición, si el lugar del estadístico fuera un número decimal, se hace interpolación entre los dos valores que ocupan las posiciones enteras, anterior y posterior.

Ejemplo:

Calcularemos el primer cuartil (Q_1), el octavo decil (D_8) y el percentil 43 (P_{43}) en el ejemplo de las estaturas de los estudiantes universitarios.

1°	2°	3°	4°	5°	6°	7°	8°	9°	10°
149	150	151	153	154	156	156	159	160	160
11°	12°	13°	14°	15°	16°	17°	18°	19°	20°
160	160	160	160	160	160	160	161	161	161
21°	22°	23°	24°	25°	26°	27°	28°	29°	30°
162	162	162	163	163	164	166	166	167	167

31°	32°	33°	34°	35°	36°	37°	38°	39°	40°
168	168	169	169	169	169	170	170	170	170
41°	42°	43°	44°	45°	46°	47°	48°	49°	50°
171	172	172	172	172	173	173	173	174	174
51°	52°	53°	54°	55°	56°	57°	58°	59°	60°
175	175	176	176	178	178	179	179	182	184

Primer cuartil (Q_1)

La posición del primer cuartil es $°Q_1 = 1.(n+1)/4 = 1.(60+1)/4 = 15{,}25°$

Como el valor 15,25 no existe, se realiza interpolación lineal entre los valores correspondientes a las posiciones 15° y 16°:

		Posición		Valor		
		15°	→	160		
1	0,25	15,25°	→	Q_1	x	0
		16°	→	160		

En realidad, en este caso no hace falta realizar los cálculos de interpolación ya que los valores coinciden. Luego, el primer cuartil toma el valor 160 cm.

Q_1 = 160 cm

Interpretación: *El 25% de las estaturas de los estudiantes universitarios observados son inferiores o iguales a 160 cm y el 75% restante son mayores o iguales a 160 cm.*

Octavo decil (D_8)

La posición del octavo decil es $°D_8 = 8.(n+1)/10 = 8.(60+1)/10 = 48{,}8°$

Como el valor 48,8 no existe, se realiza interpolación lineal entre los valores correspondientes a las posiciones 48° y 49°:

		Posición		Valor		
		48°	→	173		
1	0,8	48,8°	→	D_8	x	1
		49°	→	174		

$$\frac{0{,}8}{1} = \frac{x}{1} \quad \Rightarrow \quad x = 0{,}8$$

Luego, el octavo decil toma el valor 173,8 cm.

D_8 = 173,80 cm

Interpretación: *El 80% de las estaturas de los estudiantes universitarios observados son inferiores o iguales a 173,80 cm y el 20% restante son mayores o iguales a 173,80 cm.*

Percentil 43 (P_{43})

La posición del percentil 43 es °P_{43} = 43.(n+1)/100 = 43.(60+1)/100 = 26,23°
Como el valor 26,23 no existe, se realiza interpolación lineal entre los valores correspondientes a las posiciones 26° y 27°:

		Posición		Valor		
		26°	→	164		
1	0,23	26,23°	→	P_{43}	x	2
		27°	→	166		

$$\frac{0,23}{1} = \frac{x}{2} \Rightarrow x = 0,46$$

Luego, el percentil 43 es el valor 164,46 cm.
P_{43} = 164,46 cm
Interpretación: *El 43% de las estaturas de los estudiantes universitarios observados son inferiores o iguales a 164,46 cm y el 57% restante son mayores o iguales a 164,46 cm.*

Datos agrupados

Para calcular los estadísticos de orden en datos agrupados vamos a seguir los siguientes pasos:

- Calcular el orden o posición del estadístico que se desea conocer, con la misma fórmula usada para datos individuales.
- Buscar el valor obtenido como orden del estadístico en la columna de frecuencia acumulada (F_i), si no está, tomar el inmediato superior y llamar a la clase correspondiente *clase del cuartil*, *clase del decil* o *clase del percentil*, según corresponda.
- Diremos que el estadístico de orden pertenece a este intervalo, pero es necesaria una mayor precisión. Por esto buscaremos el valor del estadístico dentro de la clase que lo contiene.
- El valor de los estadísticos se obtiene mediante las siguientes fórmulas:

$$Q_k = L_{inf\ Qk} + l \cdot \left(\frac{\frac{k.n}{4} - F_{ant\ Q_k}}{f_{Q_k}} \right)$$

Siendo:
$L_{inf\ Qk}$: límite inferior de la clase del cuartil k.
$F_{ant\ Qk}$: frecuencia acumulada correspondiente a la clase anterior a la clase del cuartil k.
f_{Qk} : frecuencia absoluta correspondiente a la clase del cuartil k.
l : longitud de la clase del cuartil k.
n : tamaño de la muestra.

$$D_k = L_{inf\ Dk} + l \cdot \left(\frac{\frac{k.n}{10} - F_{ant\ Dk}}{f_{Dk}} \right)$$

Siendo:
$L_{inf\ Dk}$: límite inferior de la clase del decil k.
$F_{ant\ Dk}$: frecuencia acumulada correspondiente a la clase anterior a la clase del decil k.
f_{Dk} : frecuencia absoluta correspondiente a la clase del decil k.
l : longitud de la clase del decil k.
n : tamaño de la muestra.

$$P_k = L_{inf\ Pk} + l \cdot \left(\frac{\frac{k.n}{100} - F_{ant\ Pk}}{f_{Pk}} \right)$$

Siendo:
$L_{inf\ Pk}$: límite inferior de la clase del percentil k.
$F_{ant\ Pk}$: frecuencia acumulada correspondiente a la clase anterior a la clase del percentil k.
f_{Pk} : frecuencia absoluta correspondiente a la clase del percentil k.
l : longitud de la clase del percentil k.
n : tamaño de la muestra.

Ejemplo:

A modo de ejemplo, calcularemos el tercer cuartil (Q_3), el segundo decil (D_2) y el percentil 95 (P_{95}) en la serie de datos correspondiente a las estaturas de los alumnos universitarios.

Intervalos	x_i	f_i	F_i	
[149 , 154)	151,5	4	4	
[154 , 159)	156,5	3	7	
[159 , 164)	161,5	18	25	⇦ Clase del D_2
[164 , 169)	166,5	7	32	
[169 , 174)	171,5	16	48	⇦ Clase del Q_3
[174 , 179)	176,5	8	56	
[179 , 184]	181,5	4	60	⇦ Clase del P_{95}
		n=60		

Tercer cuartil (Q_3)

La posición de la clase del tercer cuartil es
$°Q_3 = 3.(n+1)/4 = 3.(60+1)/4 = 45{,}75°$.
Luego, buscando el valor obtenido en la columna de la frecuencia acumulada, se ve el intervalo correspondiente a la clase del tercer cuartil.
Como el valor obtenido en $°Q_3$ no existe, se toma el intervalo inmediato superior.

La clase del tercer cuartil es [169 ; 174). Así, identificada la clase, calculamos el valor del tercer cuartil dentro del intervalo, que se halla mediante la fórmula:

$$Q_3 = L_{inf\ Q3} + I \cdot \left(\frac{\frac{3.n}{4} - F_{ant\ Q_3}}{f_{Q_3}} \right) = 169 + 5 \cdot \left(\frac{45 - 32}{16} \right) = \mathbf{173{,}06\ cm}$$

Interpretación: *El 75% de las estaturas de los estudiantes universitarios observados son iguales o inferiores a 173,06 cm y el otro 25% son iguales o superiores a 173,06 cm.*

Segundo decil (D_2)

La posición de la clase del segundo decil es
$°D_2 = 2.(n+1)/10 = 2.(60+1)/10 = 12{,}2°$.
Luego, buscando el valor obtenido en la columna de la frecuencia acumulada, se ve el intervalo correspondiente a la clase del segundo decil.
Como el valor obtenido en $°D_2$ no existe, se toma el intervalo inmediato superior.
La clase del segundo decil es [159 ; 164). Así, identificada la clase, calculamos el valor del segundo decil dentro del intervalo, que se halla mediante la fórmula:

$$D_2 = L_{inf\ D2} + I \cdot \left(\frac{\frac{2.n}{10} - F_{ant\ D_2}}{f_{D_2}} \right) = 159 + 5 \cdot \left(\frac{12 - 7}{18} \right) = \mathbf{160{,}39\ cm}$$

Interpretación: *El 20% de las estaturas de los estudiantes universitarios observados son iguales o inferiores a 160,39 cm y el otro 80% son iguales o superiores a 160,39 cm.*

Percentil noventa y cinco (P_{95})

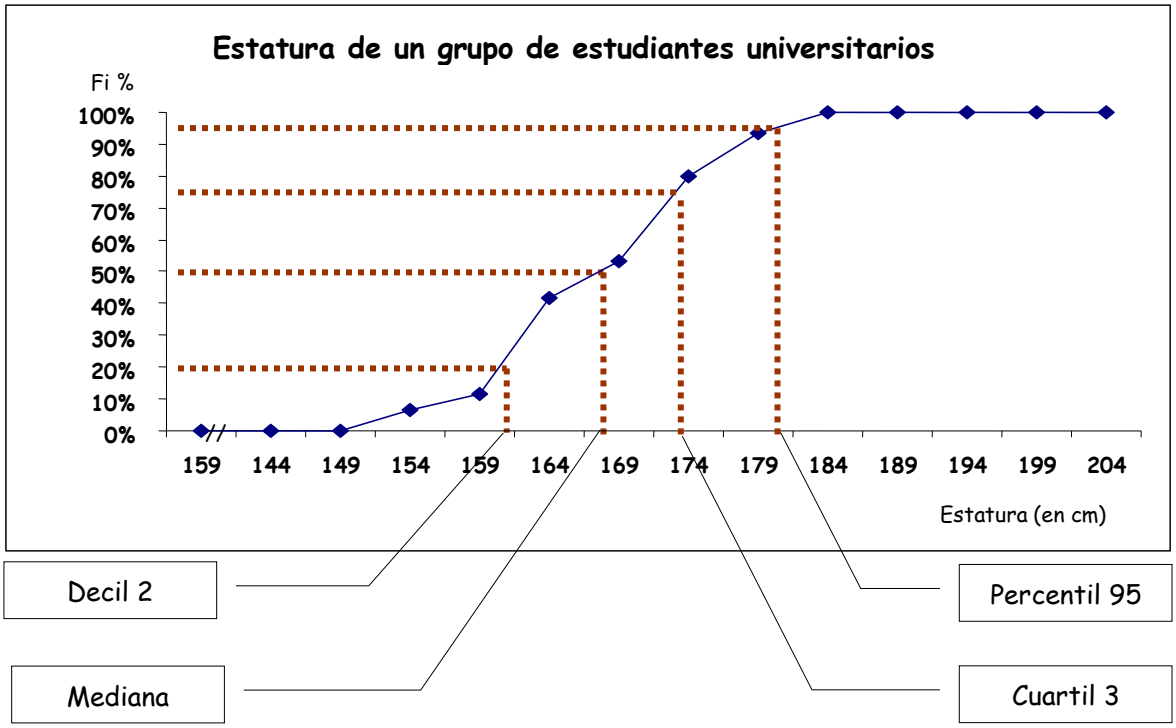
La posición de la clase del percentil noventa y cinco es
$°P_{95} = 95.(n+1)/100 = 95.(60+1)/100 = 57{,}95°$.
Luego, buscando el valor obtenido en la columna de la frecuencia acumulada, se ve el intervalo correspondiente a la clase del percentil noventa y cinco.
Como el valor obtenido en $°P_{95}$ no existe, se toma el intervalo inmediato superior.
La clase del percentil noventa y cinco es [179 ; 184]. Así, identificada la clase, calculamos el valor del percentil noventa y cinco dentro del intervalo, que se halla mediante la fórmula:

$$P_{95} = L_{inf\ P95} + I \cdot \left(\frac{\frac{95.n}{100} - F_{ant\ P_{95}}}{f_{P_{95}}} \right) = 179 + 5 \cdot \left(\frac{57 - 56}{4} \right) = \mathbf{180{,}25\ cm}$$

Interpretación: *El 95% de las estaturas de los estudiantes universitarios observados son iguales o inferiores a 180,25 cm y el otro 5% son iguales o superiores a 180,25 cm.*

Las medidas de posición no centrada pueden calcularse a partir del gráfico de la distribución acumulada (ojiva), aunque de manera aproximada.

Es conveniente realizar la ojiva colocando en ordenadas la frecuencia acumulada porcentual. Ubicar el porcentaje deseado en el eje de ordenadas y ver a qué valor de abscisa corresponde.



Ejercicio integrador

Dada la siguiente tabla, correspondiente a las edades de un grupo de personas:

Edad	Cantidad de personas
25	1
26	2
27	3
28	4
29	6
30	5
31	3
32	1

Calcular e interpretar:

a) La media aritmética
b) La mediana
c) El modo
d) El primer cuartil
e) El cuarto decil
f) El percentil 82
g) La varianza y la desviación estándar
h) El coeficiente de variación

Realizar el gráfico de puntos correspondiente.

✍ Vamos a resolverlo juntos...

Dada la siguiente tabla, correspondiente a las edades de un grupo de personas:

Edad	Cantidad de personas	*Frecuencia acumulada*
x_i	f_i	F_i
25	1	*1*
26	2	*3*
27	3	*6*
28	4	*10*
29	6	*16*
30	5	*21*
31	3	*24*
32	1	*25*

Calcular e interpretar:

a) La media aritmética

$\overline{x}$ = ***28,76 años***.

La edad promedio en este grupo de personas es, aproximadamente, de 29 años.

b) La mediana

°Me = (n+1) / 2 = 26 / 2 = 13° ⇒ ***Me = 29 años*** *(En este caso no hace falta interpolar porque la mediana está exactamente en el 13° lugar, o sea, corresponde a 29 años). Cuando se trabaja con los datos agrupados de esta manera (no en intervalos), basta con buscar la posición de la mediana (en nuestro caso, 13°) en la columna de frecuencias acumuladas y, si no está el valor buscado, elegimos el valor de la variable correspondiente a la frecuencia acumulada inmediatamente superior a la buscada (en nuestro caso, el valor es 26, correspondiente a F_i = 16).*

Esto indica que el 50% de las personas tienen 29 años o menos y el otro 50% de las personas tienen 29 años o más.

c) El modo

Mo = 29 años

Esta edad es la más frecuente porque se presentó seis veces.

d) El primer cuartil

°Q_1 = (n+1)/4 = 26/4 = 6,25° ⇒ ***Q_1 = 27,25 años*** *(Valor interpolado entre el 6° y 7° valor del conjunto de datos ordenados, o sea, entre los valores 27 y 28 años)*

Esto indica que el 25% de las personas tienen 27,25 años o menos y el otro 75% de las personas tienen 27,25 años o más.

e) El cuarto decil

°D_4 = 4.(n+1)/10 = 4.26/10 = 10,4° ⇒ ***D_4 = 28,40 años*** *(Valor interpolado entre el 10° y 11° valor del conjunto de datos ordenados, o sea, entre los valores 28 y 29 años)*

Esto indica que el 40% de las personas tienen 28,40 años o menos y el otro 60% de las personas tienen 28,40 años o más.

f) El percentil 82

°P_{82} = 82.(n+1)/100 = 82.26/100 = 21,32° ⇒ ***P_{82} = 30,32 años*** *(Valor interpolado entre el 21° y 22° valor del conjunto de datos ordenados, o sea, entre los valores 30 y 31 años)*

Esto indica que el 82% de las personas tienen 30,32 años o menos y el otro 18% de las personas tienen 30,32 años o más.

g) La varianza y la desviación estándar

s^2 = 3,106666... años^2

s = 1,762573876 años

En promedio, la edad de este grupo de personas se aparta de la media en aproximadamente 1,76 años.

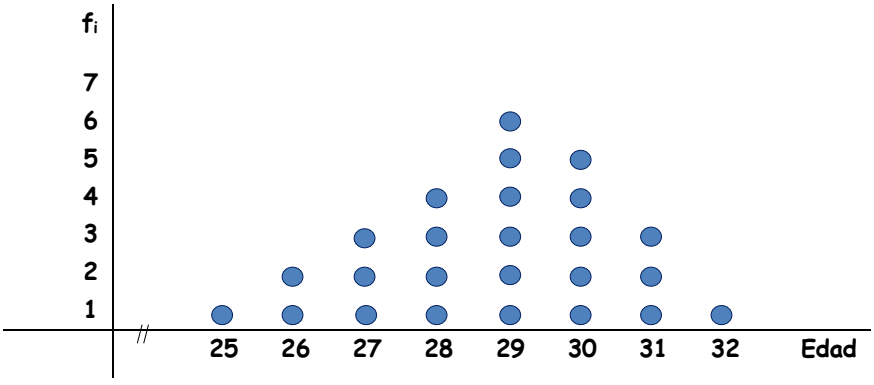
h) El coeficiente de variación

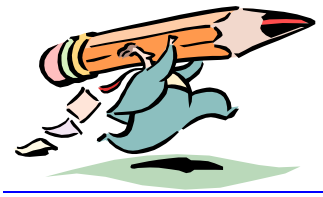
$$CV = \frac{s}{\bar{x}} = \frac{1,7626}{28,76} = 0,0613$$

La desviación estándar representa el 6,13% de la media.

Realizar el gráfico de puntos correspondiente.

Edades de un grupo de personas





A trabajar solos...

... Aunque no tanto porque al final encontrará el ejercicio resuelto.

La precipitación anual de lluvias, en centímetros, para un período de 30 años está dada como sigue:

42,3	35,7	47,5	31,2	28,3	37,0	41,3	29,3	32,4	41,3	34,3	35,2	43,0	36,3	35,7
41,5	43,2	30,7	38,4	46,5	43,2	31,7	36,8	43,6	45,2	32,8	30,7	36,2	34,7	35,3

a) Clasificar los datos y construir una tabla de distribución de frecuencias.
b) Calcular la media, la mediana, el modo, el cuartil 1, el decil 4, el percentil 86 y la desviación estándar. Interpretar los resultados obtenidos.
c) Representar gráficamente los datos en un histograma de frecuencias.

1.4 Descripción de datos: Gráfico de caja y extensiones

El gráfico de caja y extensiones fue descrito por Tukey, denominándolo 'box and whiskers'. Para su construcción se utilizan cinco medidas descriptivas de la distribución de frecuencias: el valor mínimo, el valor máximo, el primer cuartil, la mediana, el tercer cuartil y la media aritmética.

Explicaremos su construcción paso a paso:

- Antes de comenzar la representación gráfica debemos calcular algunos valores que serán necesarios para realizar el gráfico:
  - Valor mínimo: $x_{mín}$
  - Valor máximo: $x_{máx}$
  - Media aritmética: $\bar{x}$
  - Mediana: $\tilde{x}$
  - Primer cuartil: Q_1
  - Tercer cuartil: Q_3
  - Rango intercuartílico: $RI = Q_3 - Q_1$
  - $Ref1 = Q_1 - 3.RI$
  - $Ref2 = Q_1 - 1{,}5.RI$
  - $Ref3 = Q_3 + 1{,}5.RI$
  - $Ref4 = Q_3 + 3.RI$

- Se traza una línea horizontal de longitud proporcional al recorrido de la variable, que llamaremos eje. Sobre el eje se señalarán las subdivisiones que se consideren necesarias, para representar los datos de la muestra.
- Paralelamente al eje se construye una **caja** rectangular con altura arbitraria y cuya base abarca desde el primer cuartil hasta el tercer cuartil. Como vemos, esta caja indica gráficamente el intervalo de variación de *al menos* el 50% de los valores centrales de la distribución.
- La caja se divide en dos partes (no necesariamente congruentes), trazando una línea a la altura de la mediana. Cada una de estas partes indica el intervalo de variabilidad de al menos una cuarta parte de los datos.
- Luego se representan sobre el eje las referencias calculadas (Ref1, Ref2, Ref3 y Ref4) y, aunque en el gráfico definitivo no deben aparecer, se pueden colocar líneas punteadas perpendiculares al eje para distinguir claramente las referencias.
- A la caja, así dibujada, se añaden dos guías paralelas al eje, que llamaremos ***extensiones*** o ***bigotes***, una de cada lado, de la siguiente forma:
  - el *primero* de estos segmentos se prolonga, hacia la izquierda, desde el primer cuartil (o sea, desde la caja) hasta el mínimo de la distribución o hasta el valor (observado en la muestra) igual o inmediato superior a la Ref2, según cuál de estos valores sea *mayor*.
  - el *segundo* de estos segmentos se prolonga, hacia la derecha, desde el tercer cuartil (o sea, desde la caja) hasta el máximo de la distribución o hasta el valor (observado en la muestra) igual o inmediato inferior a la Ref3, según cuál de estos valores sea *menor*.
- Los valores observados en la muestra que quedan fuera del intervalo cubierto por la caja y estas extensiones, se llaman **valores apartados** ('outliers').
  - Los datos que se encuentran entre las Ref1 y Ref2 o entre las Ref3 y Ref 4 son llamados *valores atípicos*, por estar alejados de los valores centrales de la distribución. Lo indicaremos con **O**.
  - Si se observan valores menores que la Ref1 o valores mayores que Ref4, son los llamados *valores anómalos*, que son valores más alejados que los atípicos de los valores centrales de la distribución. Lo indicaremos con ✱.
- Finalmente, se indica con un signo **+** el valor de la *media aritmética*.

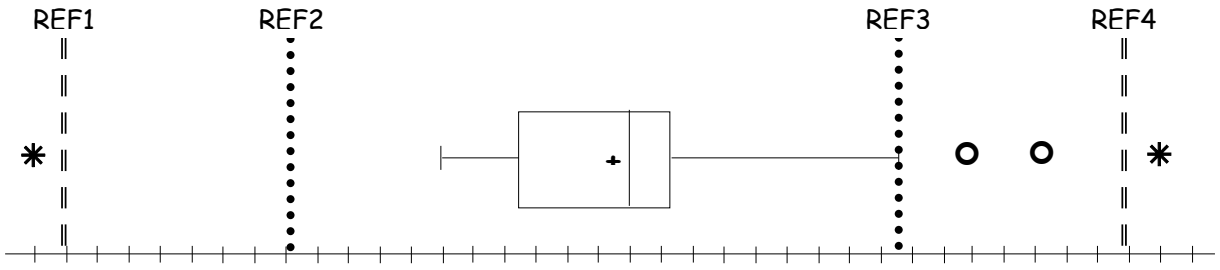
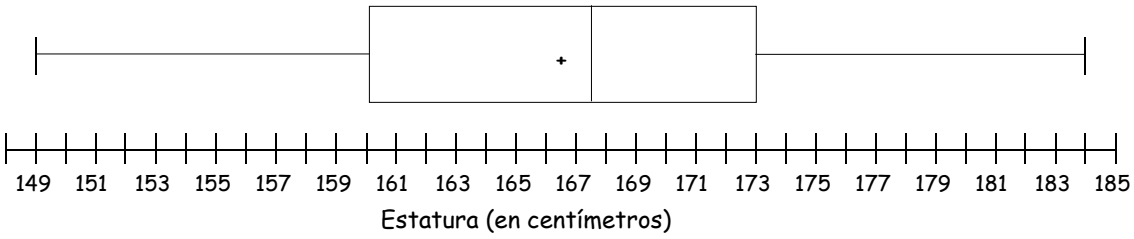


Gráfico de caja y extensiones hipotético

Ejemplo:

A partir de nuestro ejemplo (estaturas de los estudiantes universitarios), construiremos el gráfico de caja y extensiones. El gráfico de caja y extensiones se realiza sólo para datos individuales, ya que es necesario identificar, si fuera necesario, los valores atípicos.

- En primer lugar, anotaremos la información necesaria:
  - Valor mínimo: $x_{mín}$ = 149 cm
  - Valor máximo: $x_{máx}$ = 184 cm
  - Media aritmética: $\bar{x}$ = 166,55 cm
  - Mediana: $\tilde{x}$ = 167,5 cm
  - Primer cuartil: Q_1 = 160 cm
  - Tercer cuartil: Q_3 = 172,25 cm
  - Rango intercuartílico: RI = Q_3 - Q_1 = 12,25 cm
  - Ref1 = Q_1 - 3.RI = 123,25 cm
  - Ref2 = Q_1 - 1,5.RI = 141,63 cm
  - Ref3 = Q_3 + 1,5.RI = 190,63 cm
  - Ref4 = Q_3 + 3.RI = 209,00 cm
- Como el valor mínimo $x_{mín}$ = 149 cm es mayor que la Ref2 = 140,88 cm, la extensión izquierda llegará hasta 149 cm, por lo que no hay valores apartados por izquierda, es decir, no se observan valores extremadamente pequeños respecto al conjunto de datos.
- Como el valor máximo $x_{máx}$ = 184 cm es menor que la Ref3 = 191,88 cm, la extensión derecha llegará hasta 184 cm, por lo que no hay valores apartados por derecha, es decir, no se observan valores extremadamente grandes respecto al conjunto de datos.



Utilidades del gráfico de caja y extensiones

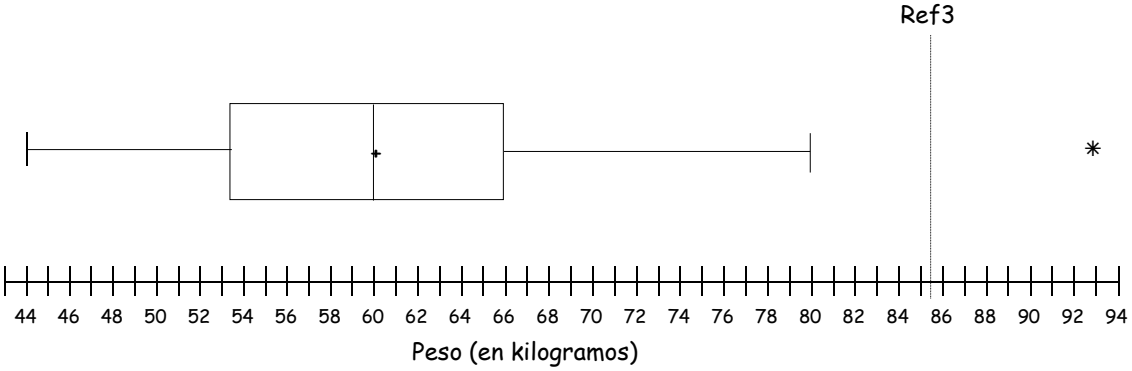
- El gráfico de caja y extensiones nos proporciona la posición relativa de la mediana, los cuartiles y extremos de una distribución.
- El gráfico de caja y extensiones nos proporciona información sobre los valores atípicos, sugiriendo la necesidad de utilizar (o no) determinados estadísticos.
- El gráfico de caja y extensiones nos informa de la simetría o asimetría de la distribución.
- El gráfico de caja y extensiones se puede utilizar para comparar la misma variable en dos muestras distintas.

Ejemplo:

Para verificar todas estas utilidades analizaremos una nueva serie estadística, que contiene los pesos, en kilogramos, de un grupo de sesenta personas:

Varones	55	64	70	74	75	70	62	93	60	62	70	71
	70	80	61	60	62	68	65	65	66	68	71	72
Mujeres	60	49	52	54	56	66	45	52	48	54	56	61
	46	50	52	53	56	68	47	50	53	57	60	64
	47	50	53	57	60	64	55	52	54	44	65	60

- En primer lugar, tomaremos la muestra en su conjunto (sin distinguir por sexo) y anotaremos la información necesaria:
  - Valor mínimo: $x_{mín}$ = 44 kg
  - Valor máximo: $x_{máx}$ = 93 kg
  - Media aritmética: $\bar{x}$ = 60,067 kg
  - Mediana: $\tilde{x}$ = 60 kg
  - Primer cuartil: Q_1 = 53 kg
  - Tercer cuartil: Q_3 = 66 kg
  - Rango intercuartílico: RI = Q_3 - Q_1 = 13 kg
  - Ref1 = Q_1 - 3.RI = 14 kg
  - Ref2 = Q_1 - 1,5.RI = 33,5 kg
  - Ref3 = Q_3 + 1,5.RI = 85,5 kg
  - Ref4 = Q_3 + 3.RI = 105 kg
- El valor mínimo $x_{mín}$ = 44 kg es mayor que la Ref2 = 33,5 kg, por lo que la extensión izquierda llega hasta 44 kg, es decir, no hay valores apartados hacia la izquierda.
- Como el valor máximo $x_{máx}$ = 93 kg es mayor que la Ref3 = 85,5 kg, la extensión derecha llega hasta el valor observado inmediato inferior a la Ref3, en nuestro caso es la valor correspondiente a 80 kg.
- Vemos, entonces, que hay un valor que supera la Ref3, por lo que habrá valores apartados. Por esto se debe analizar si es un dato atípico (entre la Ref3 y la Ref4), anómalo (mayor a la Ref4). En nuestro caso es el valor correspondiente a 93 kg, o sea, es un valor atípico en el extremo superior.



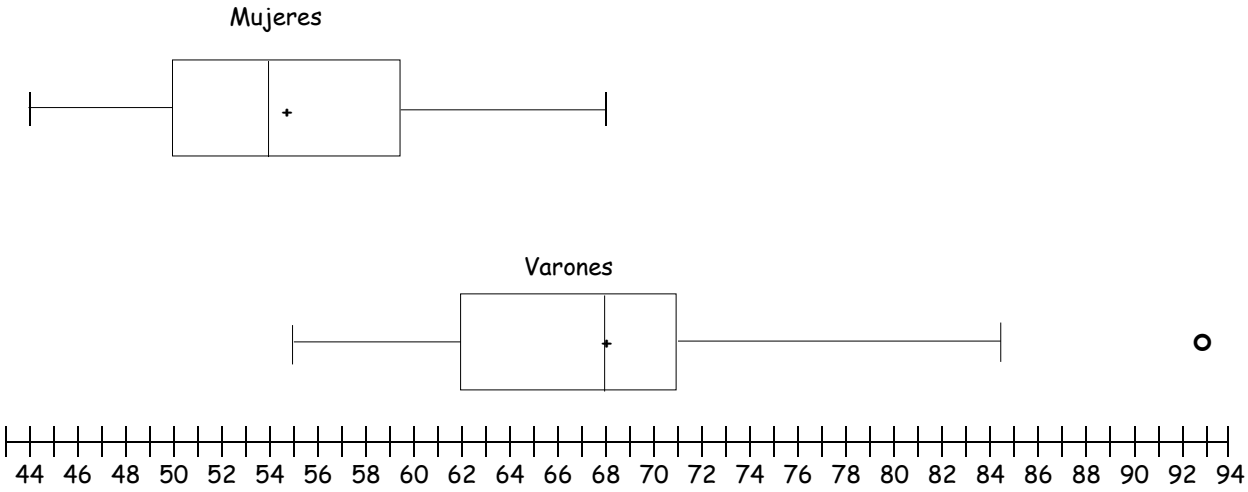
A continuación, clasificaremos la muestra según el género, realizando un gráfico de caja para cada caso, a fin de comparar ambas distribuciones:

Varones

- En primer lugar, tomaremos la muestra de los varones y anotaremos la información necesaria:
  - Valor mínimo: $x_{mín}$ = 55 kg
  - Valor máximo: $x_{máx}$ = 93 kg
  - Media aritmética: $\bar{x}$ = 68,083 kg
  - Mediana: $\tilde{x}$ = 68 kg
  - Primer cuartil: Q_1 = 62 kg
  - Tercer cuartil: Q_3 = 71 kg
  - Rango intercuartílico: RI = Q_3 - Q_1 = 9 kg
  - Ref1 = Q_1 - 3.RI = 35 kg
  - Ref2 = Q_1 - 1,5.RI = 48,5 kg
  - Ref3 = Q_3 + 1,5.RI = 84,5 kg
  - Ref4 = Q_3 + 3.RI = 98 kg
- Como $x_{mín}$ = 55 kg es mayor que la Ref2 = 48,5 kg, la extensión izquierda llegará hasta el valor mínimo.
- El $x_{máx}$ = 93 kg es mayor que la Ref3 = 84,5 kg, lo cual indica que existe por lo menos un valor apartado, por lo que la extensión derecha llegará hasta el valor inmediato anterior a la Ref3, que en nuestro ejemplo es 80 kg.
- Por lo anterior, vemos que en la muestra de varones, se presenta un valor atípico (porque está entre la Ref3 y la Ref 4) en el extremo superior.

Mujeres

- Anotaremos la información necesaria de la muestra de las mujeres:
  - Valor mínimo: $x_{mín}$ = 44 kg
  - Valor máximo: $x_{máx}$ = 68 kg
  - Media aritmética: $\bar{x}$ = 54,722 kg
  - Mediana: $\tilde{x}$ = 54 kg
  - Primer cuartil: Q_1 = 50 kg
  - Tercer cuartil: Q_3 = 60 kg
  - Rango intercuartílico: RI = Q_3 - Q_1 = 10 kg
  - Ref1 = Q_1 - 3.RI = 20 kg
  - Ref2 = Q_1 - 1,5.RI = 35 kg
  - Ref3 = Q_3 + 1,5.RI = 75 kg
  - Ref4 = Q_3 + 3.RI = 90 kg
- El $x_{mín}$ = 44 kg es mayor que la Ref2 = 35 kg, o sea, la extensión izquierda llega hasta el valor mínimo.
- El $x_{máx}$ = 68 kg es menor que la Ref3 = 75 kg, o sea, la extensión derecha llega hasta el valor máximo.
- Por lo anterior, vemos que en la muestra de mujeres, no se presenta valores atípicos ni anómalos.

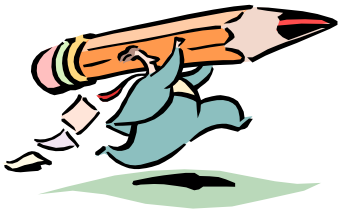


¡A repasar...!

Para autoevaluarse, responda las preguntas que están a continuación. Puede hacerlo con el material de estudio, pero asegurándose que "entiende" cada palabra, a tal punto que usted podría explicarle a un amigo, que no conoce el tema, de manera simple, los conceptos estudiados:

- ☑ **¿De qué manera se pueden presentar los datos para realizar un estudio estadístico?**
- ☑ **¿Qué ventajas y desventajas ofrece cada una de las formas de presentación de datos?**
- ☑ **¿Cómo se describe gráficamente un conjunto de datos?**
- ☑ **¿Se describen de igual manera los conjuntos de datos cualitativos que cuantitativos?**
- ☑ **¿A qué llamamos patrón de comportamiento de un conjunto de datos?**
- ☑ **¿Cómo se describe numéricamente un conjunto de datos?**
- ☑ **¿Qué medidas caracterizan a un conjunto de datos?**
- ☑ **¿Qué característica tienen las medidas de tendencia central?**
- ☑ **¿Cuáles son las ventajas y desventajas de cada una de las medidas de tendencia central?**

- ☑ **¿A qué llamamos media pesada o media ponderada?**
- ☑ **¿Qué característica tienen las medidas de dispersión?**
- ☑ **¿Qué es una puntuación Z? ¿Cuál es su utilidad?**
- ☑ **¿Qué característica tienen las medidas de posición no centradas?**
- ☑ **¿Qué aporta el gráfico de caja y extensiones al análisis de datos?**
- ☑ **¿Qué medidas descriptivas se pueden leer en un gráfico de caja y extensiones? ¿Cuáles no se pueden leer?**
- ☑ **¿Cuándo un dato es atípico y cuándo es anómalo?**



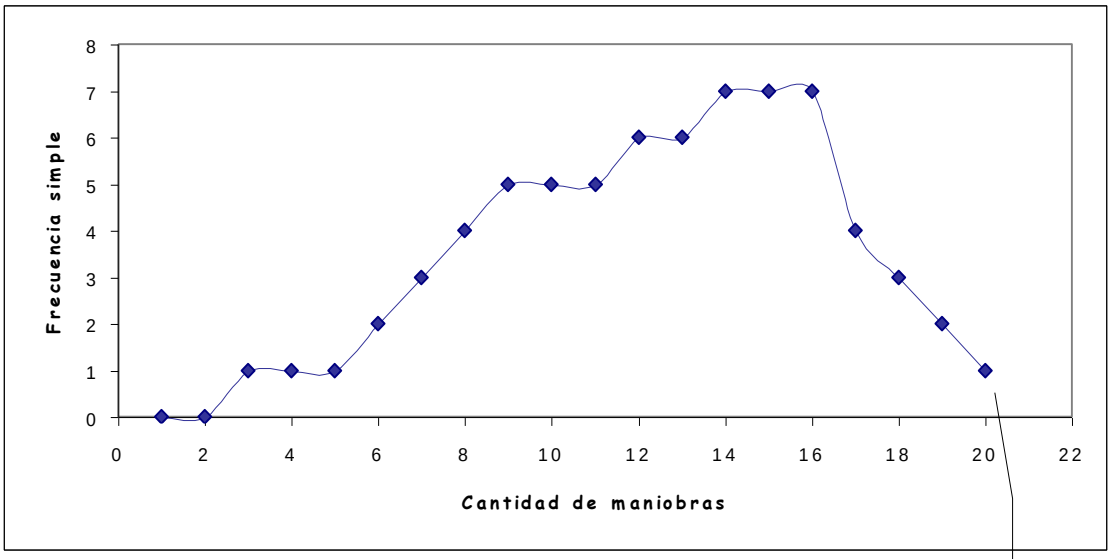
Por favor, no avance al siguiente tema si tiene dudas o no recuerda las nociones aquí volcadas. Pero si se siente listo para continuar, es hora de empezar a trabajar con las **autoevaluaciones** y las **aplicaciones prácticas**...

Respuestas

Para pensar

- Según el mito popular, ¿qué tipo de distribución tiene la variable: "Cantidad de maniobras que debe hacer una mujer para estacionar correctamente un auto, entre otros dos"?

  La respuesta depende de cuán machista o feminista sea el que conteste, pero tratando de encontrar un punto de equilibrio, y en base al mito popular, podemos decir que una gráfica más o menos representativa sería:

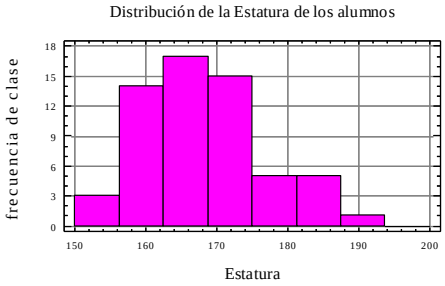


Que la gráfica termine aquí no significa que la variable no pueda tomar valores mayores a 20, sino que es en estos casos cuando las mujeres deciden pagar una playa de estacionamiento.

Nota a pedido del profesor titular (que es varón): Él no intervino en la realización de este gráfico...

- A continuación se presentan tablas y gráficos que representan el comportamiento de algunas variables analizadas en el mismo grupo de estudiantes.
  - En base a la observación de las tablas y gráficas, responda:
    - ¿A qué nivel educativo supone que pertenecen estos alumnos?

      Al observar la distribución de las estaturas se podría decir que son alumnos universitarios o de los últimos años del nivel medio.

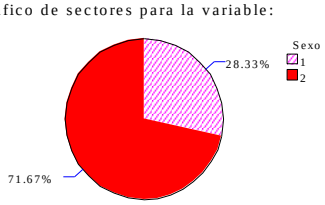


- ¿Qué tipos de chistes causarían más efecto, los machistas o los feministas?

  Según la distribución de los sexos, hay más mujeres que hombres, por lo que, probablemente, causarían más efecto los chistes feministas.

Distribución de frecuencias del SEXO de los alumnos

Sexo	Valor	Frecuencia Absoluta	Frecuencia Relativa	Acumulada Absoluta	Acumulada Relativa
Hombre	1	17	0.2833	17	0.2833
Mujer	2	43	0.7167	60	1.0000



- Respecto a la tabla DEPORTES:

  ¿Cómo definimos la variable que se refiere a la práctica deportiva?

  'Periodicidad' con que realiza actividades deportivas.

Tabla de frecuencia para la variable DEPORTE

Deporte	Valor	Frecuencia Absoluta	Frecuencia Relativa	Acumulativa Absoluta	Acumulativa Relativa
POCO	1	15	0,2500	15	0,2500
FRECUENTEMENTE	2	32	0,5333	47	0,7833
SISTEMÁTICAMENTE	3	13	0,2167	60	1,0000

POCO: Sólo de vez en cuando
FRECUENTEMENTE: Una vez por semana
SISTEMÁTICAMENTE: Dos o más veces por semana

¿Cómo la codificamos?

POCO: Sólo de vez en cuando

FRECUENTEMENTE: Una vez por semana

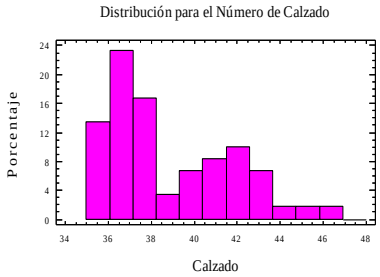
SISTEMÁTICAMENTE: Dos o más veces por semana

¿Cuál es la escala de medición?

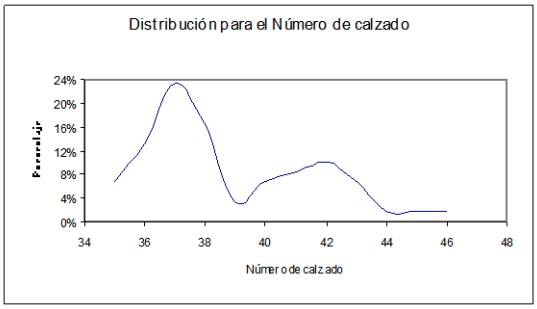
La variable es cualitativa, medida en escala ordinal.

- ¿Qué puede decir respecto al patrón de comportamiento de la variable "Número de Calzado"?

  La variable "Número de Calzado" presenta un comportamiento muy interesante, ya que la curva parece tener dos partes.



  La primera que se observa concentra los valores más pequeños, hasta el número 39 y, la segunda, concentra al resto de los valores, presentándose así, como dos "lomas".



  Este patrón de comportamiento indica que hay dos grupos claramente separados por género, donde las mujeres tienen menor número de calzado y los varones calzan más.

  ¿Cuál es la escala de medición?

  La variable es cuantitativa, medida en escala de intervalo.

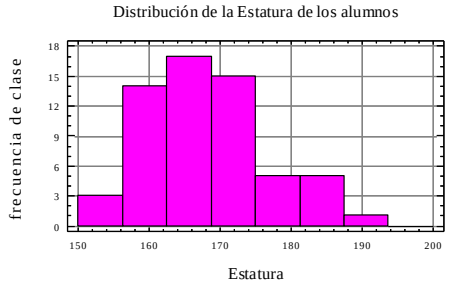
- ¿Es coherente la distribución del número de calzados con el sexo de los estudiantes? ¿Por qué?

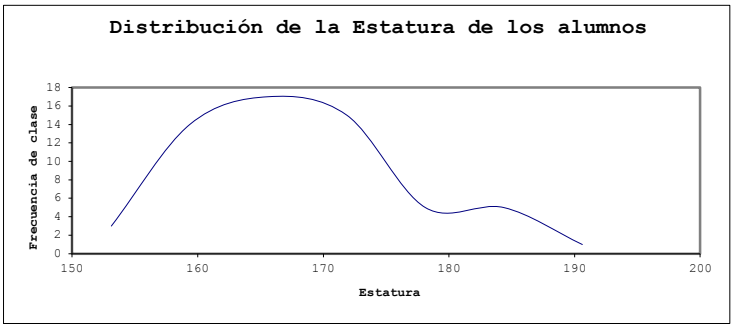
  La distribución es coherente porque logra discriminar el grupo de varones del de mujeres.

  ¿Por qué la primera "loma" es más alta? La respuesta es simple, porque hay más mujeres que varones en este grupo de estudiantes. Sólo por eso se ven con más frecuencia números bajos en el calzado.

- ¿Cómo se "comporta" la estatura de los alumnos?

  La estatura de los alumnos presenta un leve sesgo a derecha, con un intervalo modal entre 162,50 cm y 168,75 cm.





Para pensar

La siguiente es la distribución de los salarios de los empleados de una pequeña fábrica:

Salario	Cantidad de empleados
$10000	1
$2500	1
$1000	1
$500	2
$200	4

Los empleados realizan una huelga para pedir mejora de sus salarios. Un periodista realiza una nota preguntando cuál es el salario promedio.
¿Qué medida de tendencia central daría usted si...

En primer lugar vamos a calcular las medidas de tendencia central estudiadas:

- $\bar{x}$ = $ 1700
- °Me = (n+1) / 2 = 10 / 2 = 5 ⇒ Me = $ 500
- Mo = $ 200

a) ... fuera el dueño?

*Si fuera el dueño daría el valor de la **media aritmética**.*

b) ... fuera un representante sindical?

*Si fuera el representante sindical daría el valor de la **moda**.*

c) ... fuera un investigador científico?

*Si fuera un investigador científico daría el valor de la **mediana** y además aclararía que la muestra es muy heterogénea, por lo que sería necesario dar otras medidas que permitan describir mejor el conjunto de datos.*

A trabajar solos...

La precipitación anual de lluvias, en centímetros, para un período de 30 años está dada como sigue:

42,3	35,7	47,5	31,2	28,3	37,0	41,3	29,3	32,4	41,3	34,3	35,2	43,0	36,3	35,7
41,5	43,2	30,7	38,4	46,5	43,2	31,7	36,8	43,6	45,2	32,8	30,7	36,2	34,7	35,3

a) Clasificar los datos y construir una tabla de distribución de frecuencias.

28,3	29,3	30,7	30,7	31,2	31,7	32,4	32,8	34,3	34,7
35,2	35,3	35,7	35,7	36,2	36,3	36,8	37,0	38,4	41,3
41,3	41,5	42,3	43,0	43,2	43,2	43,6	45,2	46,5	47,5

$x_{mín} = 28,3$ $\quad x_{máx} = 47,5$ $\quad R = x_{máx} - x_{mín} = 19,2$

$k = 1 + 3,3.\log n \approx 5,8745 \approx 6$ $\quad l = R / k = 19,2 / 6 = 3,2$

Intervalos	x_i	f_i	F_i
[28,3 ; 31,5)	29,9	5	5
[31,5 ; 34,7)	33,1	4	9
[34,7 ; 37,9)	36,3	9	18
[37,9 ; 41,1)	39,5	1	19
[41,1 ; 44,3)	42,7	8	27
[44,3 ; 47,5]	45,9	3	30

b) Calcular la media, la mediana, el modo, el cuartil 1, el decil 4, el percentil 86 y la desviación estándar. Interpretar los resultados obtenidos.

Media aritmética:

$\overline{x} =$ **37,58 cm**

- *La precipitación anual de lluvias promedio, en los últimos 30 años, es de 37,58 cm.*

Mediana:

$^{\circ} Me = (n+1) / 2 = 31 / 2 = 15,5 \Rightarrow Me \in [34,7 ; 37,9)$

$$\mathbf{Me} = Li_{Me} + l.\left(\frac{\frac{n}{2} - F_{ant\ Me}}{f_{Me}}\right) = 34,7 + 3,2.\left(\frac{\frac{30}{2} - 9}{9}\right) = \mathbf{36,8333...\ cm}$$

- *En el 50% de los años la precipitación anual fue de 36,83 cm o menos y en el otro 50% la precipitación anual fue de 36,83 cm o más.*

Modo:

$Mo \in [34,7 ; 37,9)$ *por ser el intervalo con mayor frecuencia absoluta.*

$$\mathbf{Mo} = \mathbf{x_{Mo}} = L_{inf\ Mo} + l.\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) = 34,7 + 3,2.\left(\frac{5}{5 + 8}\right) = \mathbf{35,93\ cm}$$

Siendo $\Delta_1 = 9 - 4 = 5$ $\quad$ *y* $\Delta_2 = 9 - 1 = 8$

- *La precipitación anual más frecuente, en los últimos 30 años, es de 35,93 cm.*

Primer cuartil:

$^{o}Q_1 = (n+1) / 4 = 7{,}75 \Rightarrow \quad Q_1 \in [31{,}5 \ ; \ 34{,}7)$

$$Q_1 = L_{inf\ Q_1} + I. \left(\frac{\frac{1.n}{4} - F_{ant\ Q_1}}{f_{Q_1}} \right) = 31{,}5 + 3{,}2. \left(\frac{7{,}5 - 5}{4} \right) = \mathbf{33{,}5\ cm}$$

- *En el 25% de los años la precipitación fue de 33,5 cm o menos y el otro 75% de los años la precipitación fue de 33,5 cm o más.*

Cuarto decil:

$^{o}D_4 = 4.(n+1) / 10 = 12{,}4 \Rightarrow \quad D_4 \in [\ 34{,}7 \ ; \ 37{,}9\ )$

$$D_4 = L_{inf\ D_4} + I. \left(\frac{\frac{4.n}{10} - F_{ant\ D_4}}{f_{D_4}} \right) = 34{,}7 + 3{,}2. \left(\frac{12 - 9}{9} \right) = \mathbf{35{,}7666...\ cm}$$

- *En el 40% de los años la precipitación fue de 35,77 cm o menos y el otro 60% de los años la precipitación fue de 35,77 cm o más.*

Percentil 86:

$^{o}P_{86} = 86.(n+1)/100 = 26{,}66 \Rightarrow P_{86} \in [41{,}1 \ ; \ 44{,}3)$

$$P_{86} = L_{inf\ P_{86}} + I. \left(\frac{\frac{86.n}{100} - F_{ant\ P_{86}}}{f_{P_{86}}} \right) = 41{,}1 + 3{,}2. \left(\frac{25{,}8 - 19}{8} \right) = \mathbf{43{,}82\ cm}$$

- *En el 86% de los años la precipitación fue de 43,82 cm o menos y el otro 14% de los años la precipitación fue de 43,82 cm o más.*

Desviación estándar:

$\mathbf{s = 5{,}288269\ cm}$

- *En promedio, la precipitación anual de lluvias se aparta de la media en aproximadamente 5,2883 cm.*

c) Representar gráficamente los datos en un histograma de frecuencias.