

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, sólo incluimos los más usados:
 - o *Análisis estadístico simple*: esta parte proporciona el fundamento básico para el análisis estadístico.

- o *Inducción estadística*: analiza una población o universo basada en un estudio muestral. Otros métodos estadísticos distintos de los inductivos son referidos como Estadística Descriptiva.
- o *Análisis de series de tiempo*: analiza los cambios en las actividades de negocios y económicas.
- o *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 la toma 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 estas 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.

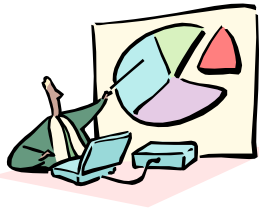


La síntesis puede ser tan simple como transmitir un hecho específico o lo bastante compleja, como para contener varias fuentes, varios formatos de presentación o diferentes medios de información y permitir la comunicación efectiva de ideas abstractas.

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

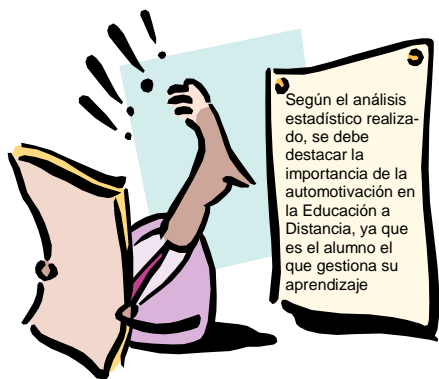
¿Se ha preguntado por qué algunas presentaciones logran captar su atención durante horas, mientras que otras pierden su atractivo en cuestión de minutos?

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 en Estadística.

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.**

Cuando se comienza a 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: *fuentes 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 1992, se analizó las cantidades (en miles de toneladas) de importaciones de acero, en distintos países:

Principales fuentes de importaciones de acero en Estados Unidos durante 1992

| Países | Frecuencia simple absoluta | Frecuencia simple relativa | Frecuencia simple relativa porcentual |
|----------------------|----------------------------|----------------------------|---------------------------------------|
| x_i | f_i | fr_i | $fr_i \%$ |
| Bélgica y Luxemburgo | 1247 | 0,3041 | 30,41 % |
| Japón | 1072 | 0,2615 | 26,15 % |
| Alemania | 460 | 0,1122 | 11,22 % |
| Canadá | 367 | 0,0895 | 8,95 % |
| Francia | 299 | 0,0729 | 7,29 % |
| Reino Unido | 250 | 0,0610 | 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 1992.

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: $fr_i \cdot 100\%$.

Las tablas estadísticas para variables cuantitativas son similares a las anteriores, aunque, en este caso, la variable puede ser ordenada 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:

Estatuta de sesenta estudiantes universitarios de Mendoza en 2004

| Valores observados | Frecuencia simple absoluta | Frecuencia simple relativa | Frecuencia simple relativa porcentual | Frecuencia acumulada absoluta | Frecuencia acumulada relativa | Frecuencia acumulada relativa porcentual |
|--------------------|----------------------------|----------------------------|---------------------------------------|-------------------------------|-------------------------------|--|
| x_i | f_i | $fr_i = f_i / n$ | $fr_i\%$ | F_i | $Fr_i = F_i/n$ | $Fr_i\%$ |
| 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 hipot ticos

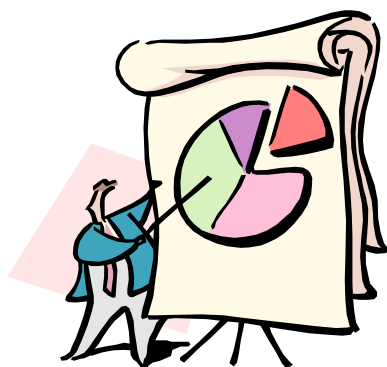
- **Variable (x_i):** para poder operar con los datos de la tabla o referirnos a ella, podemos representar la caracter stica a observar (estatura de los estudiantes universitarios) mediante la variable X y a la modalidad i - sima de dicha variable con la notaci n x_i .
- **Frecuencia simple absoluta (f_i):** representa el n mero de individuos que presentan cada modalidad x_i .
- **Frecuencia simple relativa (fr_i):** nos permite valorar la representatividad de cada categor a respecto al total de los datos. Se calcula: f_i / n .
- **Frecuencia simple relativa porcentual ($fr_i\%$):** representa en porcentajes las frecuencias relativas. Se calcula: $fr_i \cdot 100\%$.
- **Frecuencia acumulada (F_i):** representa el n mero 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 categor a 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: $Fr_i \cdot 100\%$.

Muchas veces, es necesario o resulta m s c modo trabajar con los datos agrupados en intervalos (o clases). La manera de agrupar los datos ser  estudiada m s adelante, por ahora planteamos una posibilidad de agrupaci n para ver la aplicaci n en nuestro ejemplo:

Estatura de sesenta estudiantes universitarios de Mendoza en 2004

| Intervalos o clases | Punto medio | Frecuencia simple absoluta | Frecuencia simple relativa | Frecuencia simple relativa porcentual | Frecuencia acumulada absoluta | Frecuencia acumulada relativa | Frecuencia acumulada relativa porcentual |
|---------------------------|----------------|----------------------------------|----------------------------------|--|-------------------------------------|-------------------------------------|---|
| | x_i | f_i | fr_i | $fr_i\%$ | F_i | Fr_i | $Fr_i\%$ |
| [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. **Hay que advertir, sin embargo, que 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

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 la 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**.



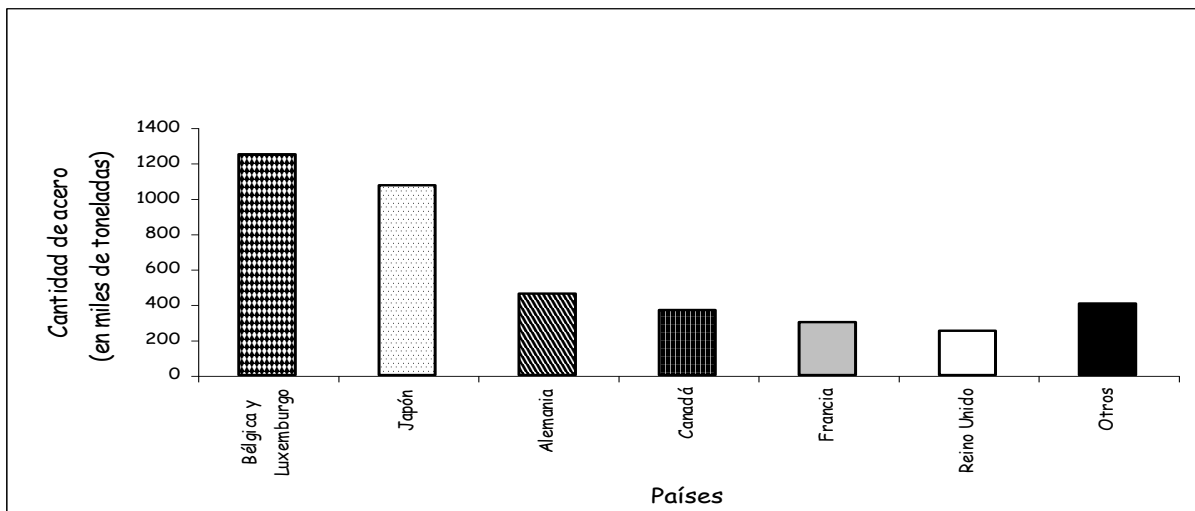
Ejemplo:

Veremos las distintas representaciones gráficas en el ejemplo anterior:



Gráfico de barras verticales

Principales fuentes de importaciones de acero en Estados Unidos durante 1992

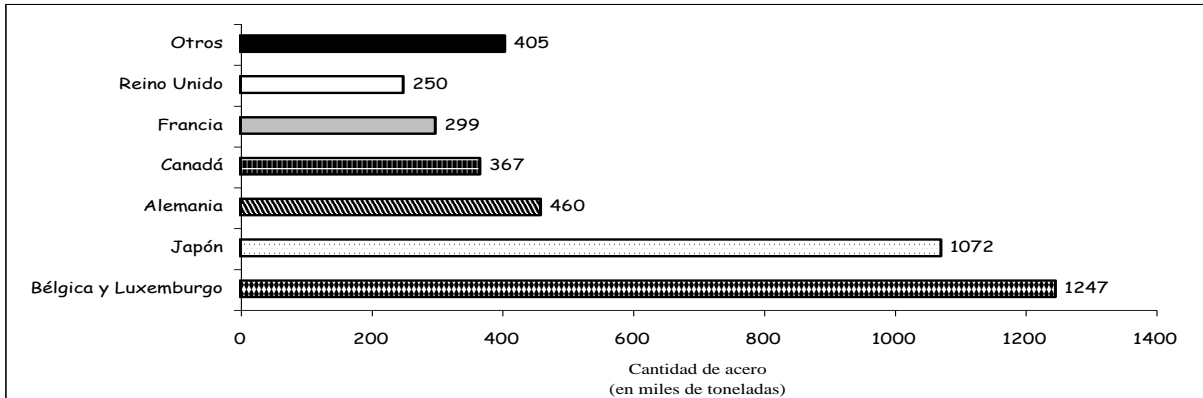


Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 1992.



Gr fico de barras horizontales

Principales fuentes de importaciones de acero en Estados Unidos durante 1992



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 1992.

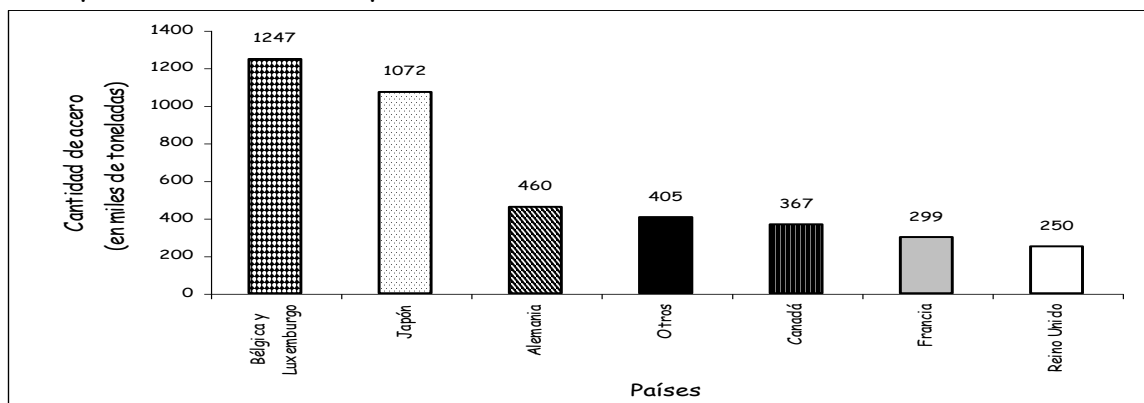


Gr fico de Pareto

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

Las categor as est n ordenadas de modo tal que en la parte izquierda aparezca la categor a con mayor frecuencia, seguida por la segunda mayor frecuencia y as , sucesivamente. Este tipo de diagramas debe su nombre al economista italiano V. Pareto.

Principales fuentes de importaciones de acero en Estados Unidos durante 1992



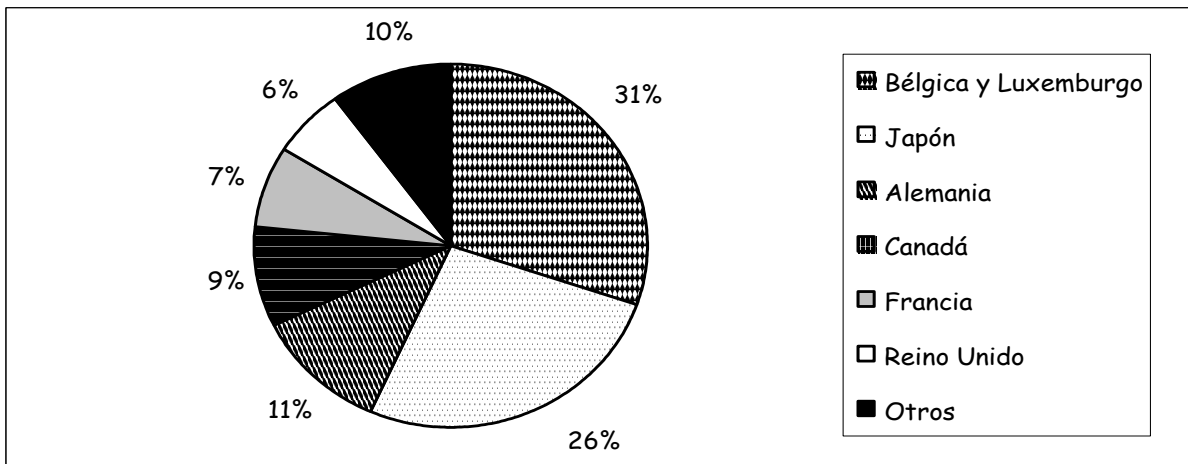
Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 1992.



Gráfico de sectores

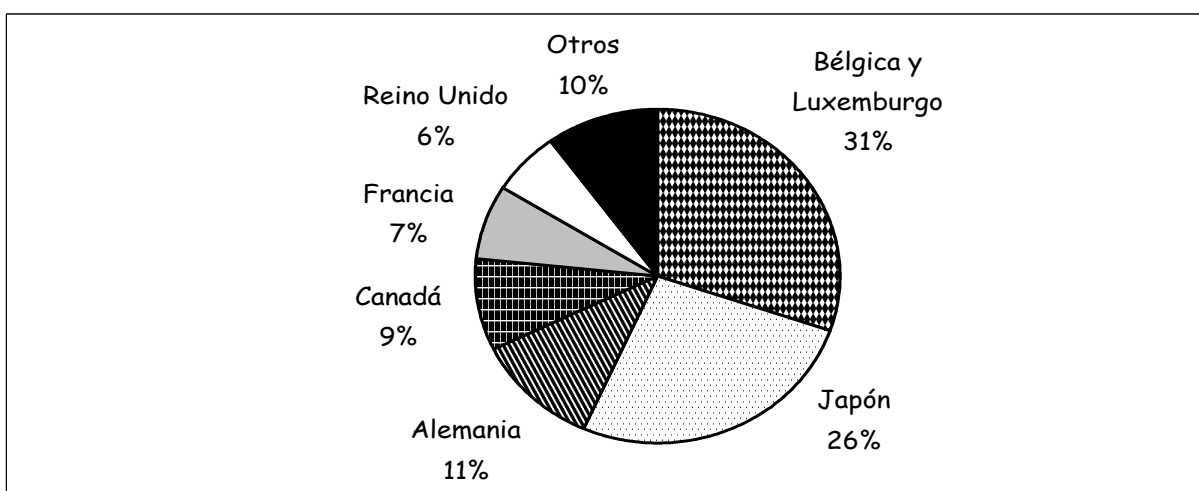
Los gráficos de sectores se utilizan para representar variables cualitativas, indicando la proporción en que cada uno de sus valores se presenta.

Principales fuentes de importaciones de acero en Estados Unidos durante 1992



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 1992.

Principales fuentes de importaciones de acero en Estados Unidos durante 1992



Fuente: U.S. Department of Commerce. Datos preparados por el American Iron and Steel Institute, publicados en Charting Steel's Progress in 1992.

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 determinado 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 (5).

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

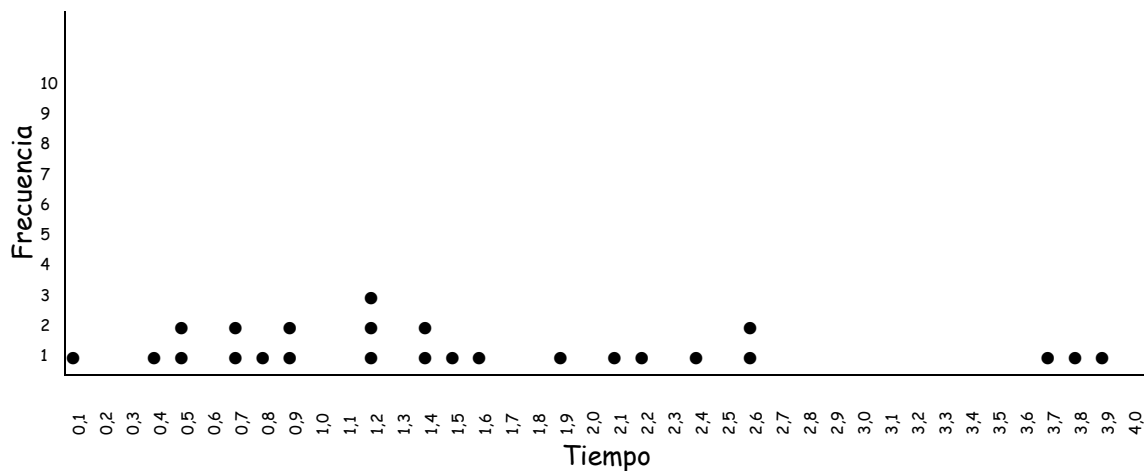
| | | |
|-----|-----------|---|
| 2 | 0 24 | Hasta el momento sólo hay 2 valores contados. |
| 9 | 0 5577899 | En este tronco hay 7 valores, más los 2 del tronco anterior, que suman 9. |
| (5) | 1 22244 | Este tronco contiene a la <i>mediana</i> , por lo que se registra la frecuencia absoluta simple 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 una serie estadística.
- **Máximo ($x_{\text{máx}}$):** se llama máximo de una variable estadística al mayor valor que toma la variable en toda la serie estadística.
- **Mínimo ($x_{\text{mín}}$):** 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 del intervalo.
- **Límite inferior de la clase (L_i):** es el mínimo valor del 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: $1 + 3,3 \cdot \log n$. 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 de textos 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 m todo 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 tama o de muestra

$n = 60$

- Reconocer el m ximo y el m nimo

$x_{\text{m x}} = 184$ $x_{\text{m n}} = 149$

- Calcular el alcance o recorrido

$R = x_{\text{m x}} - x_{\text{m n}} = 184 - 149 = 35$

- Calcular la cantidad de intervalos

$k = 1 + 3,3 \cdot \log n = 1 + 3,3 \cdot \log 60 \approx 6,87 \Rightarrow k = 7$ (El valor de k siempre debe ser redondeado a un n mero entero inferior o superior. Lo usual es hacer el redondeo matem tico.)

- Calcular la longitud de cada intervalo

$l = R / k = 35 / 7 = 5$ (Si el valor de l resultara ser un n mero 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   6,3   7, entre otras opciones, pero nunca 6,27   6,2   6.)

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

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 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 hipot ticos

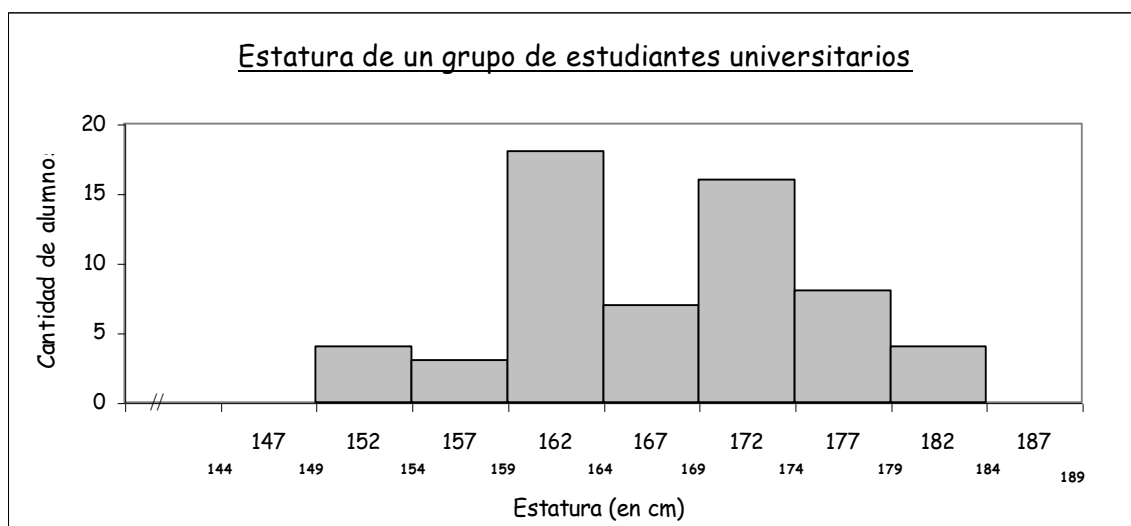
- Nota 1: Como el límite superior de cada clase coincide con el límite inferior de la siguiente, adoptamos como criterio que los intervalos se suponen semiabiertos por la derecha, es decir, en cada clase se incluyen los valores de la variable que sean mayores o iguales al **límite superior**, pero estrictamente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

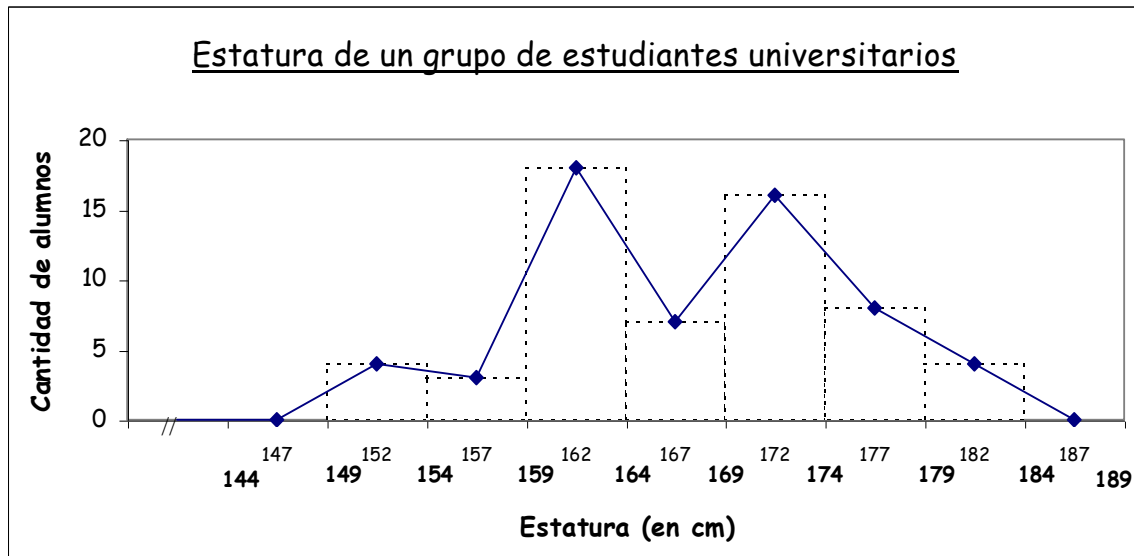
Un histograma se obtiene construyendo sobre unos ejes cartesianos rectángulos cuyas áreas son proporcionales a las frecuencias de cada intervalo. Sobre el eje de abscisas se representan dos intervalos, el anterior al primero y el posterior al último, que no tienen valores en ellos.



Fuente: Datos hipotéticos

Polígono de frecuencias

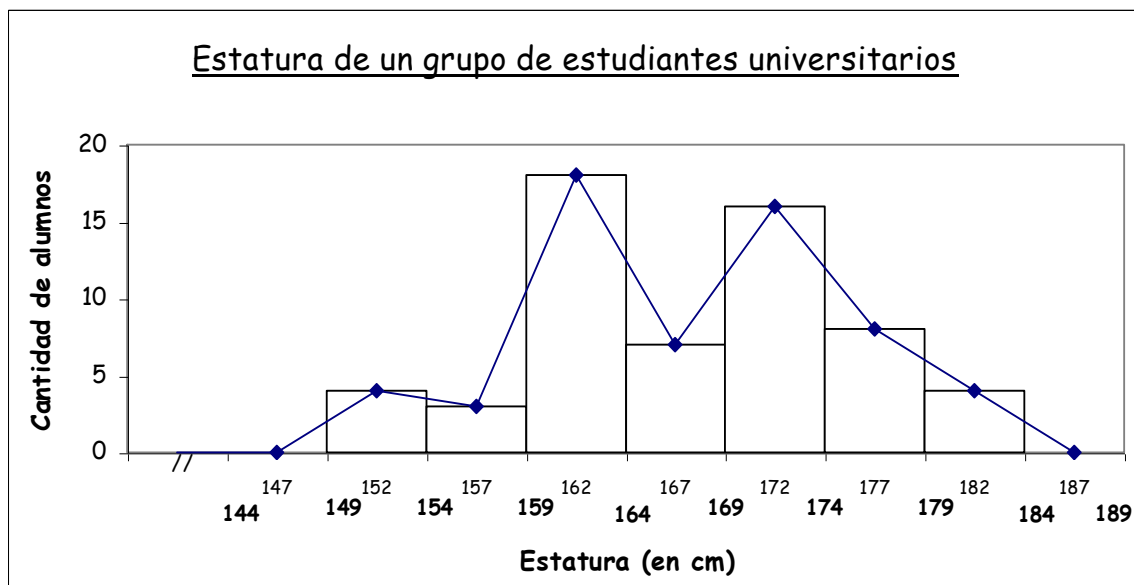
Otra forma de representar los datos es el polígono de frecuencias, que es la poligonal que resulta de unir, con segmentos, los puntos medios de las bases superiores de los rectángulos de un histograma de frecuencias. **Además, completaremos la poligonal, uniendo los puntos medios del intervalo anterior al primero de nuestra muestra y posterior al último de nuestra muestra.**



Fuente: Datos hipotéticos

Histograma y polígono de frecuencias

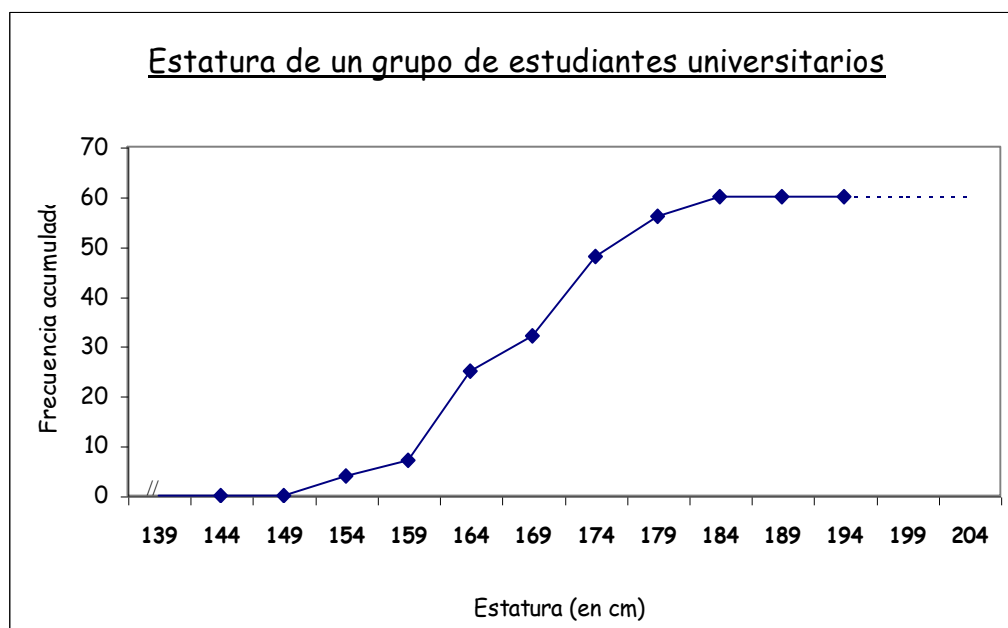
Usualmente se presentan ambos gráficos en el mismo sistema de ejes coordenados.



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

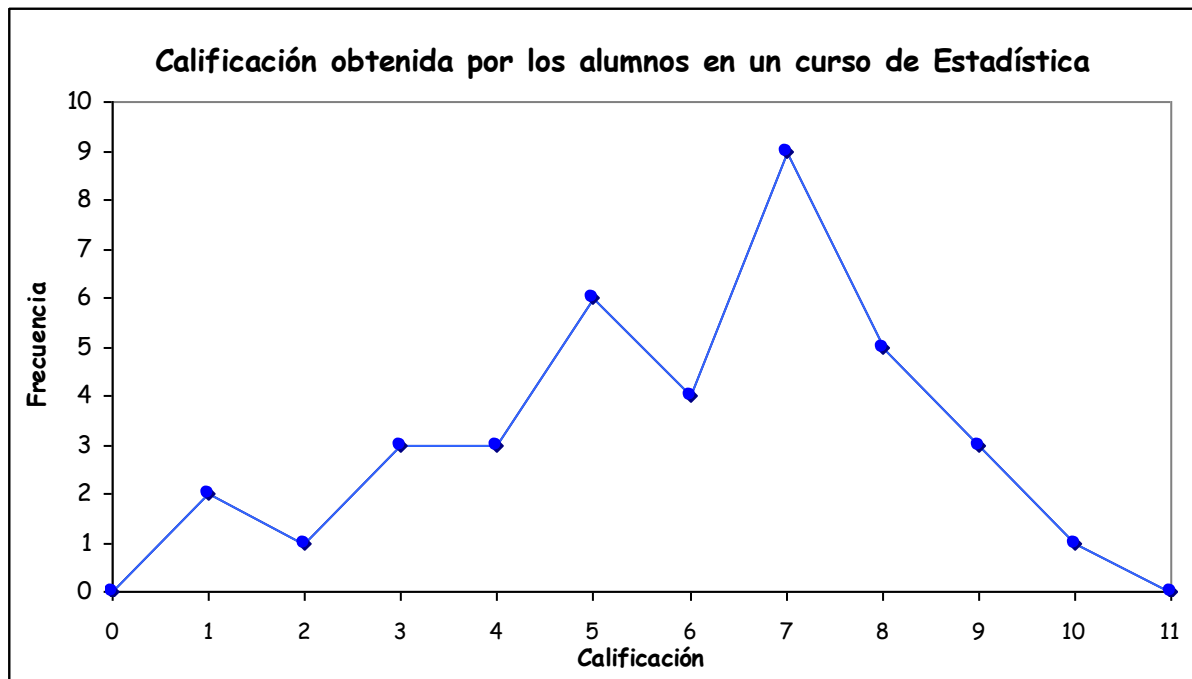
Una tabla de frecuencias, un histograma o un polígono de frecuencias describen una *distribución de frecuencias*, es decir, muestran el patrón de distribución de las frecuencias. En general, las descripciones se refieren a aspectos de la forma del histograma o del polígono de frecuencias.

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.

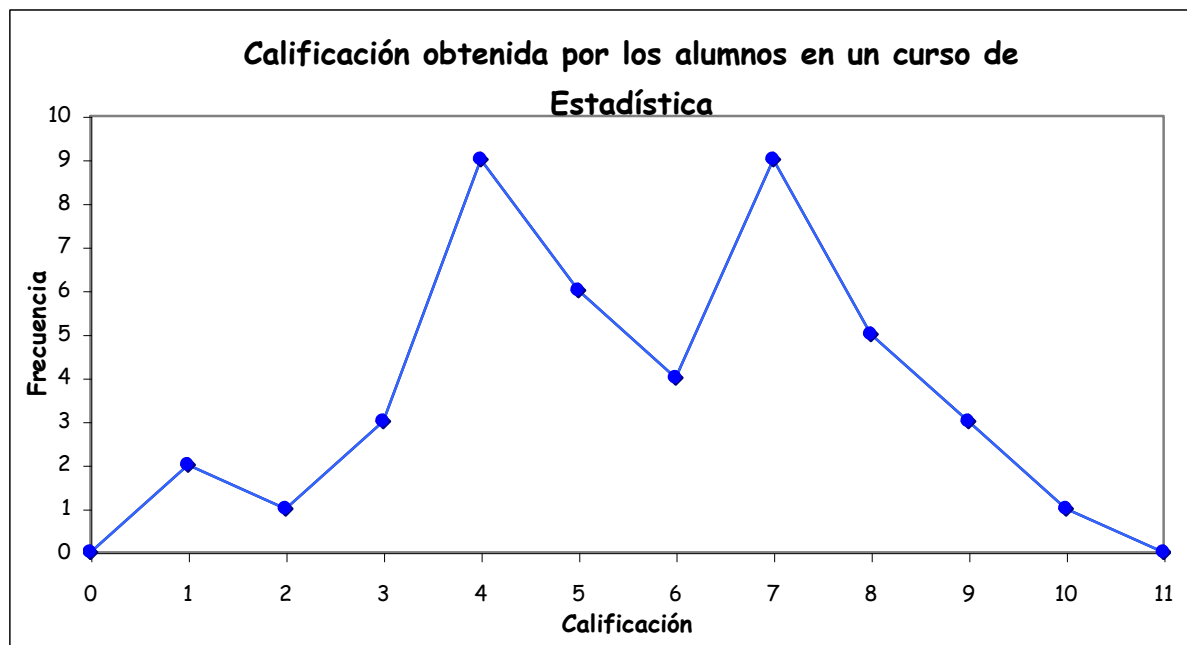
Estos valores máximos se llaman **modos** o **modas** o **valores modales**, en el caso de datos agrupados por clases, a la o las clases que presentan la máxima frecuencia se las llama **clases modales**.

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.

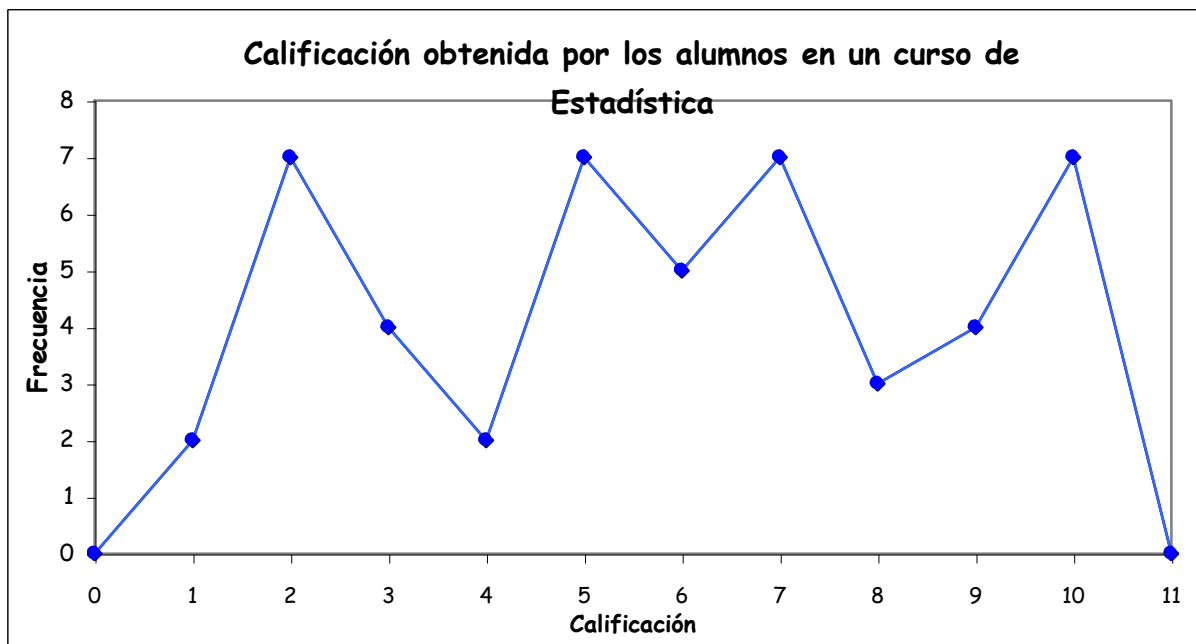
Ejemplo de distribución unimodal:



Ejemplo de distribución bimodal:



Ejemplo de distribución multimodal:



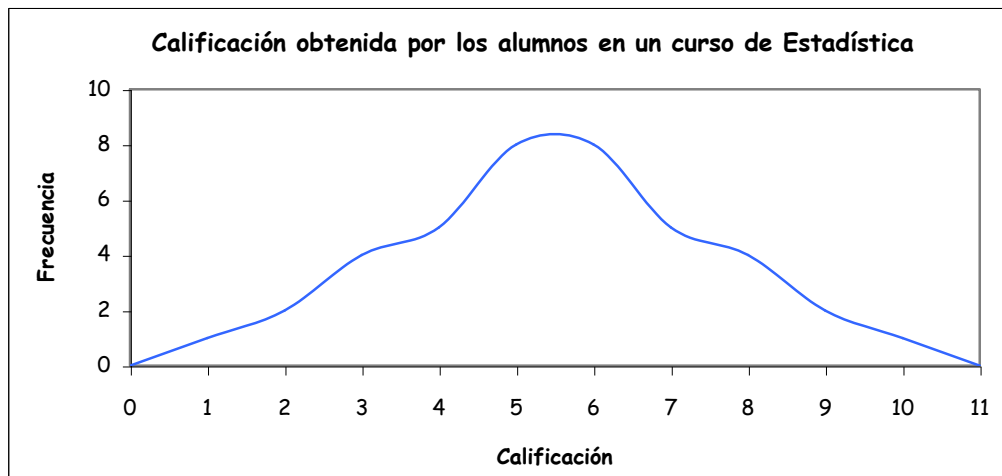
¿Y qué pasa si todos los valores presentan la misma frecuencia...?

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

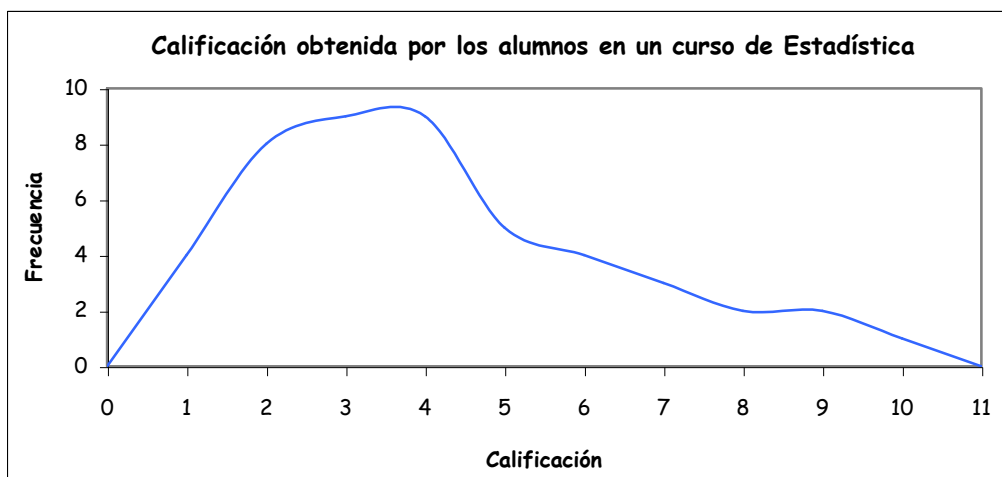
Otro aspecto a destacar, para analizar el patrón de comportamiento de un conjunto de datos, es la **simetría o asimetría** de la distribución de frecuencias.

Ejemplos de distribuciones simétricas y asimétricas con distintas asimetrías:

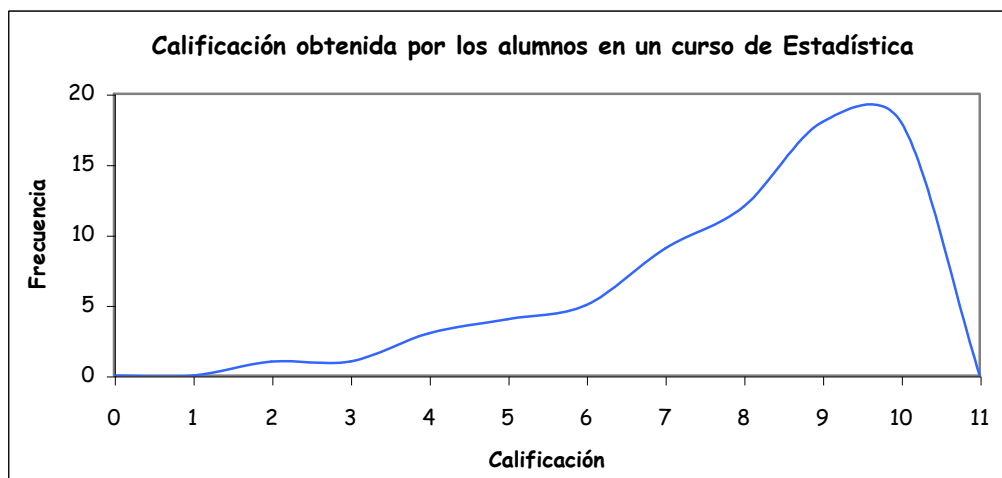
Curso A



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

| Sexo | Valor | Frecuencia | | Acumulada | |
|--------|-------|------------|----------|-----------|----------|
| | | 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:

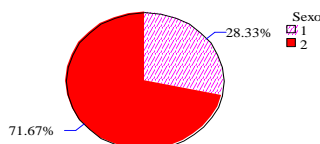


Tabla de frecuencia para la variable DEPORTE

| Deporte | Valor | Frecuencia | | Acumulativa | |
|------------------|-------|------------|----------|-------------|----------|
| | | Absoluta | Relativa | Absoluta | 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

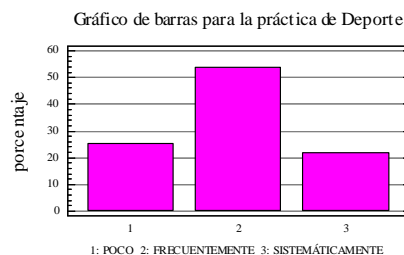


Tabla de frecuencias para PELO y OJOS

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

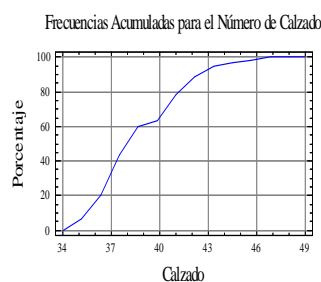
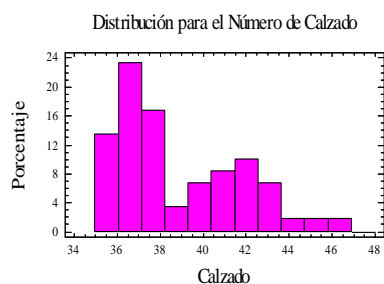
Gráfico de Tallos y Hojas para la variable Número de Calzado de los alumnos

Unidad = 0.1 35|0 representa 35.0

```

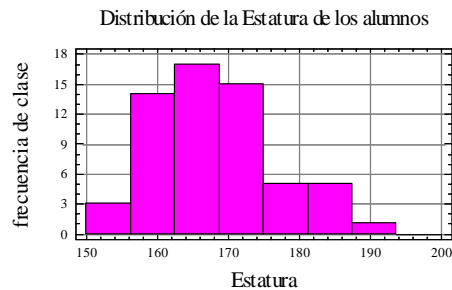
4 35|0000
12 36|00000000
26 37|00000000000000
(10) 38|0000000000
24 39|00
22 40|0000
18 41|00000
13 42|000000
7 43|0000
3 44|0
2 45|0
1 46|0

```



Distribución de la Estatura de los alumnos (cm)

| Clase | Límite de Clase Inferior | Límite de Clase Superior | Marca de Clase | Frecuencias Absoluta | Frecuencias Relativa | Acumulativa Absoluta | Acumulativa 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 es también una medida descriptiva del 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 pueden concentrar mejor la atención en varias propiedades de un conjunto de datos que se investiga.

En realidad, 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, o la velocidad con que sube y baja la distribución de izquierda a derecha, 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.

Hay varias medidas de tendencia central, con propiedades particulares que las hacen 'típicas' en alguna forma única.

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, \dots, 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. Sin embargo, 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 resto de los 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"

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 observar 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 la existencia de 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.
- ☞ Debido a que la mediana es una posición promedio, debemos ordenar los datos antes de llevar a cabo cualquier cálculo, lo cual 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 variables cualitativas nominales.**
- El valor de la moda no se ve afectada 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, nosotros 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 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, \dots, x_n es el *promedio aritmético* de éstas, se denota por:

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

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

Al dividir f_i por n , obtenemos 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 w_i , por lo que $f_i/n = w_i$. Así, y continuando el trabajo en la expresión de la media aritmética, tenemos:

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

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

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

En nuestro ejemplo, indicamos la media aritmética como:

$$\bar{x} = \frac{\sum x_i \cdot f_i}{n} = \frac{149.1 + \dots + 166.2 + \dots + 184.1}{60} = \frac{9990}{60} = 166,55 \text{ cm}$$

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 $n/2 = 60/2 = 30^\circ$ corresponde al valor 167 cm

La posición $(n+2)/2 = (60+2)/2 = 31^\circ$ corresponde al valor 168 cm

Luego, la mediana es el valor promedio entre 167 cm y 168 cm, es decir:

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

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 |
| 161 | 3 | 20 |
| . | . | . |
| . | . | . |
| . | . | . |
| 184 | 1 | 60 |
| $n = 20$ | | |

⇒ Máxima frecuencia absoluta ⇒ Valor modal

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:

$$Mo = 160 \text{ cm}$$

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.

$$\bar{x} = \frac{\sum x_i \cdot f_i}{n} = \frac{151,5 \cdot 4 + \dots + 181,5 \cdot 4}{60} = \frac{10030}{60} = 167,17 \text{ cm}$$

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 |
| [169 , 174) | 171,5 | 16 | 48 |
| [174 , 179) | 176,5 | 8 | 56 |
| [179 , 184] | 181,5 | 4 | 60 |
| | | n=60 | |

⇔ Clase mediana

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.

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

- Buscar el valor obtenido como orden de la mediana en la columna de frecuencia acumulada (F_i), 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:

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

Siendo:

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

$F_{\text{ant } \tilde{x}}$: 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.

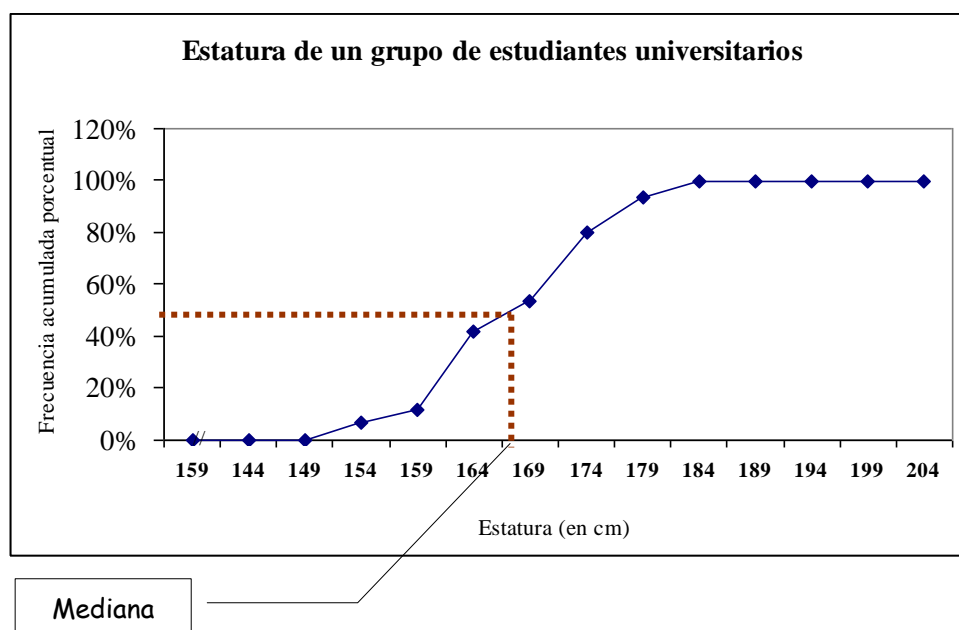
$$\tilde{x} = L_{\inf \tilde{x}} + l \cdot \left(\frac{\frac{n}{2} - F_{\text{ant } \tilde{x}}}{f_{\tilde{x}}} \right) = 164 + 5 \cdot \left(\frac{\frac{60}{2} - 25}{7} \right) = 167,57 \text{ cm}$$

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: M_d , M_e 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 |
| [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 | |

⇐ Clase modal

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, M_o , pertenece a este intervalo, pero es necesaria una mayor precisión. Por esto buscaremos el valor de la moda dentro de la clase modal.

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

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

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

Siendo:

$L_{inf \ M_o}$: 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

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

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 medio.

¿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.

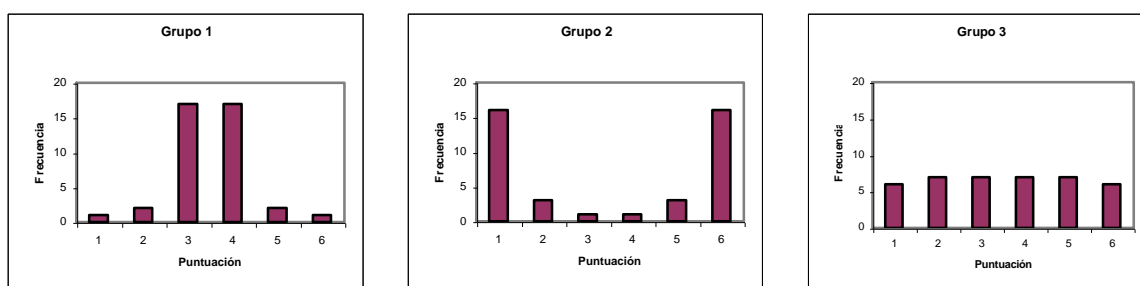
Todas ellas son valores mayores o iguales a cero, indicando un valor cero, la ausencia de dispersión.

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 | | Grupo 2 | | Grupo 3 | |
|---------|-------|---------|-------|---------|-------|
| x_i | f_i | x_i | f_i | x_i | f_i |
| 1 | 1 | 1 | 16 | 1 | 6 |
| 2 | 2 | 2 | 3 | 2 | 7 |
| 3 | 17 | 3 | 1 | 3 | 7 |
| 4 | 17 | 4 | 1 | 4 | 7 |
| 5 | 2 | 5 | 3 | 5 | 7 |
| 6 | 1 | 6 | 16 | 6 | 6 |

Puntuaciones en tres grupos de alumnos



Las tres distribuciones tienen la misma media aritmética, 2,5 puntos, ¿pero podemos afirmar que hay homogeneidad entre los grupos?. Gráficamente vemos que el valor de la media aritmética no es suficiente para describir cada una de las situaciones.

Para precisar mejor lo que denominamos como 'dispersión' podemos calcular unos 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.
$$R = x_{\text{máx}} - x_{\text{mín}}$$
- 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.
- Para muchos problemas tiene una mayor 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.

Varianza

- La **varianza** de las observaciones x_1, x_2, \dots, 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.

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

Desviación Estándar

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

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

- 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 ± 2 desviaciones estándar a partir de la media de la distribución, y al menos 89% de los valores caen dentro de ± 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 ± 1 desviación estándar a partir de la media.

- Aproximadamente 95% de los valores de la población cae dentro de ± 2 desviación estándar a partir de la media.
- Aproximadamente 99% de los valores de la población cae dentro de ± 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

$$R = x_{\max} - x_{\min} = 184 \text{ cm} - 149 \text{ cm} = 35 \text{ cm}$$

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

Varianza

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot f_i}{n - 1} = \frac{(149 - 166,55)^2 \cdot 1 + \dots + (166 - 166,55)^2 \cdot 2 + \dots + (184 - 166,55)^2 \cdot 1}{60 - 1} = 66,18 \text{ cm}^2$$

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

Desviación estándar

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

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

Coefficiente de variación

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

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

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

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

Varianza

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot f_i}{n - 1} = \frac{(151,5 - 167,17)^2 \cdot 4 + \dots + (166,5 - 167,17)^2 \cdot 7 + \dots + (181,5 - 167,17)^2 \cdot 4}{60 - 1} = 63,11 \text{ cm}^2$$

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

Desviación estándar

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

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

Coefficiente de variación

$$CV = \frac{s}{\bar{x}} = \frac{7,94 \text{ cm}}{167,17 \text{ cm}} = 0,0475$$

$$CV\% = 4,75\%$$

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.

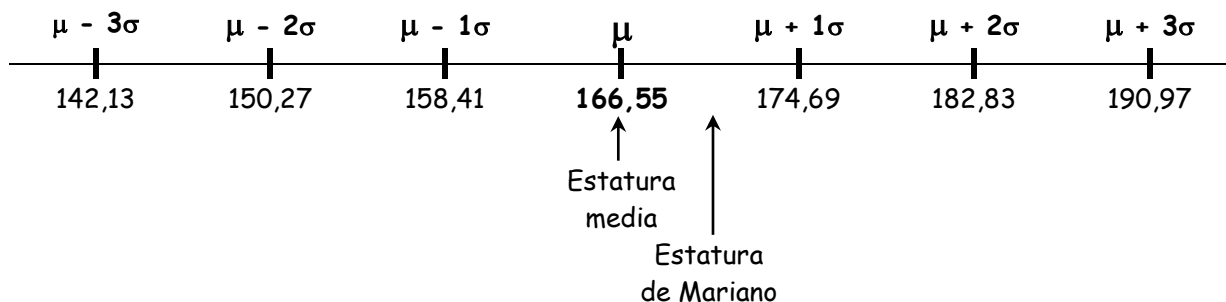


Ejemplo:

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 $\mu = 166,55$ cm y la desviación estándar como $\sigma = 8,14$ cm.



¿Qué es una puntuación Z?

Una puntuación Z es la transformación de una observación que describe mejor el lugar que esa observación ocupa 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.

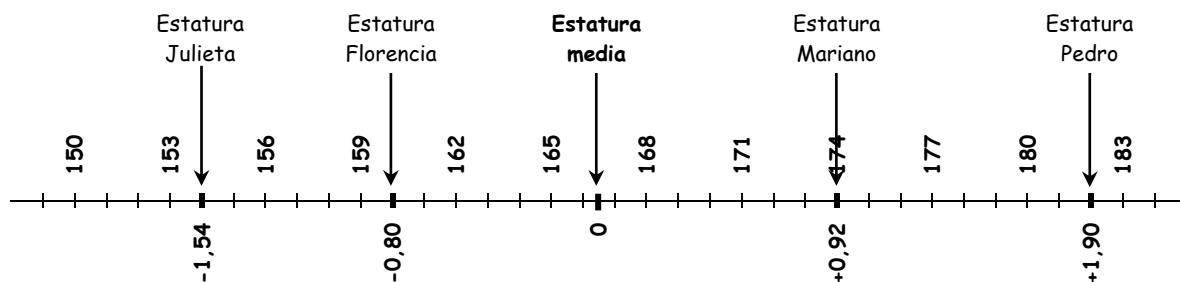
Llamaremos **puntuación bruta** al valor observado, antes de ser convertido en una puntuación Z.



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:
 $Z = (174 \text{ cm} - 166,55 \text{ cm}) / 8,14 = +0,92$.
- Florencia, que mide **160 cm**, tiene una puntuación Z:
 $Z = (160 \text{ cm} - 166,55 \text{ cm}) / 8,14 = -0,80$.

- Pedro, que mide **182 cm**, tiene una puntuación Z:
 $Z = (182 \text{ cm} - 166,55 \text{ cm})/8,14 = +1,90$.
- Julieta, que mide **154 cm**, tiene una puntuación Z:
 $Z = (154 \text{ cm} - 166,55 \text{ cm})/8,14 = -1,54$.

¿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 $X = Z \cdot \sigma + \mu$

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 de variables completamente diferentes en puntuaciones Z, podemos compararlas 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, \dots, 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, \dots, 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

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:

A modo de 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 $^{\circ}Q_1 = 1.(n+1)/4 = 1.(60+1)/4 = 15,25^{\circ}$

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

| | Posición | | Valor | |
|-------------------|-----------------|---------------|-------|----------------|
| | 15° | \rightarrow | 160 | |
| $1 \downarrow$ | $15,25^{\circ}$ | \rightarrow | Q_1 | \downarrow |
| $0,25 \downarrow$ | 16° | \rightarrow | 160 | $x \downarrow$ |
| | | | | 0 |

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: Significa que 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 $^{\circ}D_8 = 8.(n+1)/10 = 8.(60+1)/10 = 48,8^{\circ}$

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

| | Posición | | Valor | |
|------------------|----------------|---------------|-------|----------------|
| | 48° | \rightarrow | 173 | |
| $1 \downarrow$ | $48,8^{\circ}$ | \rightarrow | D_8 | \downarrow |
| $0,8 \downarrow$ | 49° | \rightarrow | 174 | $x \downarrow$ |
| | | | | 1 |

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

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

$D_8 = 173,80$ cm

Interpretación: Significa que 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 $^{\circ}P_{43} = 43.(n+1)/100 = 43.(60+1)/100 = 26,23^{\circ}$

Como el valor 26,23 no existe, se realiza interpolación lineal entre los valores correspondientes a las posiciones 26° y 27° :

| | Posición | Valor |
|--------|----------|-------------------|
| 1 ↓ | 26° | → 164 |
| 0,23 ↓ | 26,23° | → P ₄₃ |
| | 27° | → 166 |
| | | ↓ x ↓ 2 |

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

Luego, el percentil 43 es el valor 164,46 cm.

P₄₃ = 164,46 cm

Interpretación: Significa que 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* calculado en cada caso.
- 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 obtienen mediante las fórmulas:**

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

Siendo:

$L_{\inf Q_k}$: límite inferior de la clase del cuartil k.

$F_{\text{ant} Q_k}$: frecuencia acumulada correspondiente a la clase anterior a la clase del cuartil k.

f_{Q_k} : frecuencia absoluta correspondiente a la clase del cuartil k.

l : longitud de la clase del cuartil k.

n : tamaño de la muestra.

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

Siendo:

$L_{\inf D_k}$: límite inferior de la clase del decil k .

$F_{\text{ant} D_k}$: frecuencia acumulada correspondiente a la clase anterior a la clase del decil k .

f_{D_k} : frecuencia absoluta correspondiente a la clase del decil k .

l : longitud de la clase del decil k .

n : tamaño de la muestra.

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

Siendo:

$L_{\inf P_k}$: límite inferior de la clase del percentil k .

$F_{\text{ant} P_k}$: frecuencia acumulada correspondiente a la clase anterior a la clase del percentil k .

f_{P_k} : 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 |
| [164 , 169) | 166,5 | 7 | 32 |
| [169 , 174) | 171,5 | 16 | 48 |
| [174 , 179) | 176,5 | 8 | 56 |
| [179 , 184] | 181,5 | 4 | 60 |
| | | $n=60$ | |

⇔ Clase del D_2

⇔ Clase del Q_3

⇔ Clase del P_{95}

Tercer cuartil (Q_3)

La posición de la clase del tercer cuartil es

$$^{\circ}Q_3 = 3.(n+1)/4 = 3.(60+1)/4 = 45,75^{\circ}.$$

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 número obtenido en $^{\circ}Q_3$ no existe, se toma el intervalo inmediato superior.

La clase del tercer cuartil es [169 ; 174).

Una vez identificada la clase del tercer cuartil, calculamos el valor del tercer cuartil dentro del intervalo, que se halla mediante la fórmula:

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

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

$$^{\circ}D_2 = 2.(n+1)/10 = 2.(60+1)/10 = 12,2^{\circ}.$$

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 número obtenido en $^{\circ}D_2$ no existe, se toma el intervalo inmediato superior.

La clase del segundo decil es [159 ; 164).

Una vez identificada la clase del segundo decil, calculamos el valor del segundo decil dentro del intervalo, que se halla mediante la fórmula:

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

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

$$^{\circ}P_{95} = 95.(n+1)/100 = 95.(60+1)/100 = 57,95^{\circ}.$$

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 n mero obtenido en $^oP_{95}$ no existe, se toma el intervalo inmediato superior.

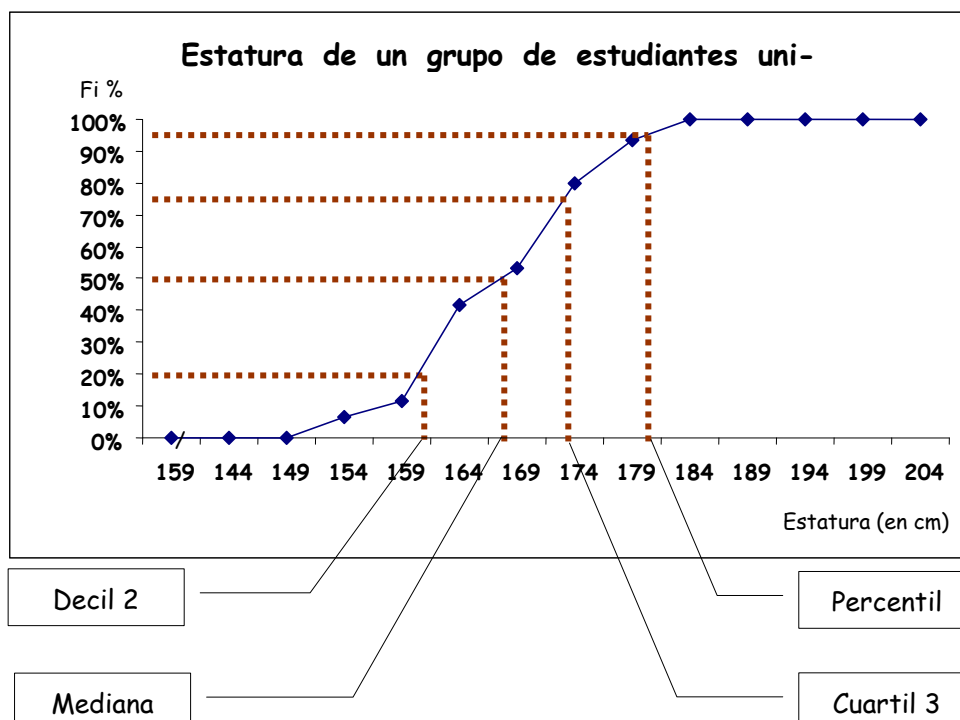
La clase del percentil noventa y cinco es [179 ; 184].

Una vez identificada la clase del percentil noventa y cinco, calculamos el valor del percentil noventa y cinco dentro del intervalo, que se halla mediante la f rmula:

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

Interpretaci n: 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.

Nota: Las medidas de posici n no centrada pueden calcularse a partir del gr fico de la distribuci n 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 qu  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:

- La media aritmética
- La mediana
- El modo
- El primer cuartil
- El cuarto decil
- El percentil 82
- La varianza y la desviación estándar
- El coeficiente de variación

Realizar el polígono de frecuencia 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

$$\bar{x} = 28,76 \text{ años.}$$

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

b) La mediana

$$^{\circ}Me = (n+1) / 2 = 26 / 2 = 13 \Rightarrow Me = 29 \text{ años (En este caso no hace falta interpolar porque la mediana está exactamente en el 13º lugar, o sea, corresponde a 29 años)}$$

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

$$Mo = 29 \text{ años}$$

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

d) El primer cuartil

$$^{\circ}Q_1 = (n+1)/4 = 26/4 = 6,25^{\circ} \Rightarrow Q_1 = 27,25 \text{ años (Valor interpolado entre el 6º y 7º valor del conjunto de datos ordenados, o sea, entre los valores 27 y 28 años)}$$

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

$$^{\circ}D_4 = 4.(n+1)/10 = 4.26/10 = 10,4^{\circ} \Rightarrow D_4 = 28,40 \text{ años (Valor interpolado entre el 10º y 11º valor del conjunto de datos ordenados, o sea, entre los valores 28 y 29 años)}$$

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

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

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

$$s^2 = 3,106666... \text{ años}^2$$

$$s = 1,762573876 \text{ años}$$

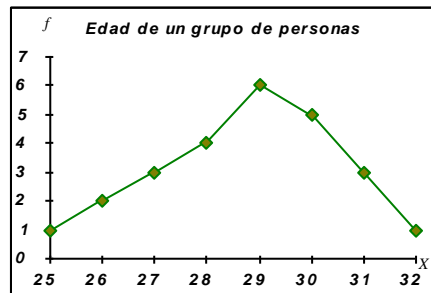
En promedio la edad de este grupo de personas se aparta de la media en aproximadamente 1,76 años.

h) El coeficiente de variación

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

El desvío estándar representa el 6,13% de la media.

Realizar el polígono de frecuencia correspondiente.



A trabajar solos...

Aunque no tanto porque al final encontrará el ejercicio resuelto.

La precipitación anual de lluvias, aproximada a décima de centímetro, para un período de 30 años es 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

- Clasificar los datos y construir una tabla de distribución de frecuencias.
- Calcular la media, la mediana, el modo, el cuartil 1, el decil 4, el percentil 86 y la desviación estándar. Interpretar los resultados obtenidos.
- 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'.

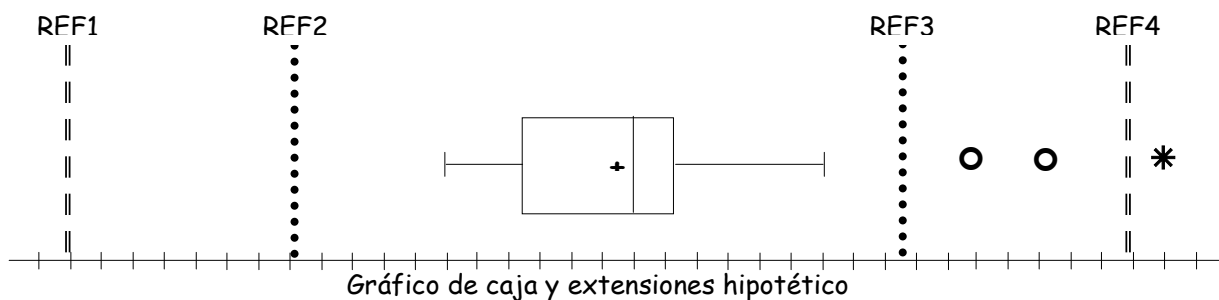
Para su construcción se utilizan cinco estadísticos de la distribución de frecuencias: el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo.

Explicaremos su construcción paso a paso:

- Antes de comenzar la graficación debemos calcular algunos valores que serán necesarios para realizar el gráfico:
 - o Valor mínimo: x_{\min}
 - o Valor máximo: x_{\max}
 - o Mediana: \tilde{x}
 - o Media aritmética: \bar{x}
 - o Primer cuartil: Q_1
 - o Tercer cuartil: Q_3
 - o Rango intercuartílico: $RI = Q_3 - Q_1$
 - o $REF1 = Q_1 - 3.RI$
 - o $REF2 = Q_1 - 1,5.RI$
 - o $REF3 = Q_3 + 1,5.RI$
 - o $REF4 = Q_3 + 3.RI$
- 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, trazando una línea a la altura de la mediana. Cada una de estas partes indica, pues, el intervalo de variabilidad de al menos una cuarta parte de los datos.
- 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:
 - o el *primero* de estos segmentos se prolonga, hacia la izquierda, desde el primer cuartil (o sea, desde la caja) hasta el valor (observado en la muestra) igual o inmediato superior a la REF2.
 - o el *segundo* de estos segmentos se prolonga, hacia la derecha, desde el tercer cuartil (o sea, desde la caja) hasta el valor (observado en la muestra) igual o inmediato inferior a la REF3.
- Si alguno de los valores observados en la muestra queda fuera del intervalo cubierto por la caja y estas extensiones, se señala en el gráfico mediante un

símbolo que lo represente como un punto. Estos datos son los llamados *valores apartados* ('outliers'),

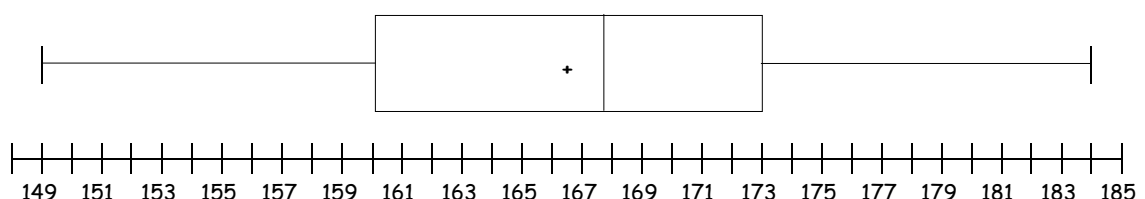
- En particular, los valores que se encuentran entre la REF 1 y la REF 2 o entre la REF 3 y la REF 4 son llamados *valores atípicos*. Los indicaremos con ○.
- En particular, los valores menores que la REF 1 o mayores que la REF 4 son llamados *valores anómalos*. Los indicaremos con *.
- Finalmente, se indica con un signo + el valor de la *media aritmética*.



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: $x_{\min} = 149$ cm
 - Valor máximo: $x_{\max} = 184$ cm
 - Mediana: $\tilde{X} = 167,5$ cm
 - Media aritmética: $\bar{X} = 166,55$ cm
 - Primer cuartil: $Q_1 = 160$ cm
 - Tercer cuartil: $Q_3 = 172,75$ cm
 - Rango intercuartílico: $RI = Q_3 - Q_1 = 12,75$ cm
 - $REF1 = Q_1 - 3 \cdot RI = 121,75$ cm
 - $REF2 = Q_1 - 1,5 \cdot RI = 140,88$ cm
 - $REF3 = Q_3 + 1,5 \cdot RI = 191,88$ cm
 - $REF4 = Q_3 + 3 \cdot RI = 211,00$ cm
- La extensión izquierda llega hasta el valor *149 cm*. La extensión derecha llega hasta el valor *184 cm*. Vemos que en nuestra muestra no se presentan valores atípicos ni valores anómalos.



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.

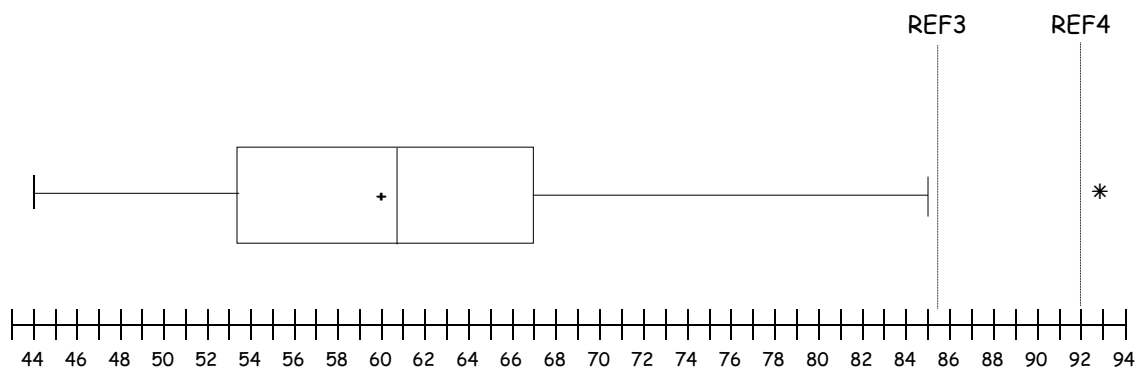


Ejemplo:

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 | 85 | 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:
 - o Valor mínimo: $x_{\min} = 44$ kg
 - o Valor máximo: $x_{\max} = 93$ kg
 - o Mediana: $\tilde{x} = 60$ kg
 - o Media aritmética: $\bar{x} = 60,15$ kg
 - o Primer cuartil: $Q_1 = 53$ kg
 - o Tercer cuartil: $Q_3 = 66$ kg
 - o Rango intercuartílico: $RI = Q_3 - Q_1 = 13$ kg
 - o $REF1 = Q_1 - 3 \cdot RI = 14$ kg
 - o $REF2 = Q_1 - 1,5 \cdot RI = 33,5$ kg
 - o $REF3 = Q_3 + 1,5 \cdot RI = 85,5$ kg
 - o $REF4 = Q_3 + 3 \cdot RI = 92$ kg
- Debemos elegir el valor observado inmediato superior a la $REF2 = 33,5$ kg para saber hasta dónde llegará la extensión izquierda.
- Debemos elegir el valor observado inmediato inferior a la $REF3 = 85,5$ kg para saber hasta dónde llegará la extensión derecha.
- Por lo anterior, vemos que en nuestra muestra, se presenta un valor anómalo en el extremo superior de la muestra.



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

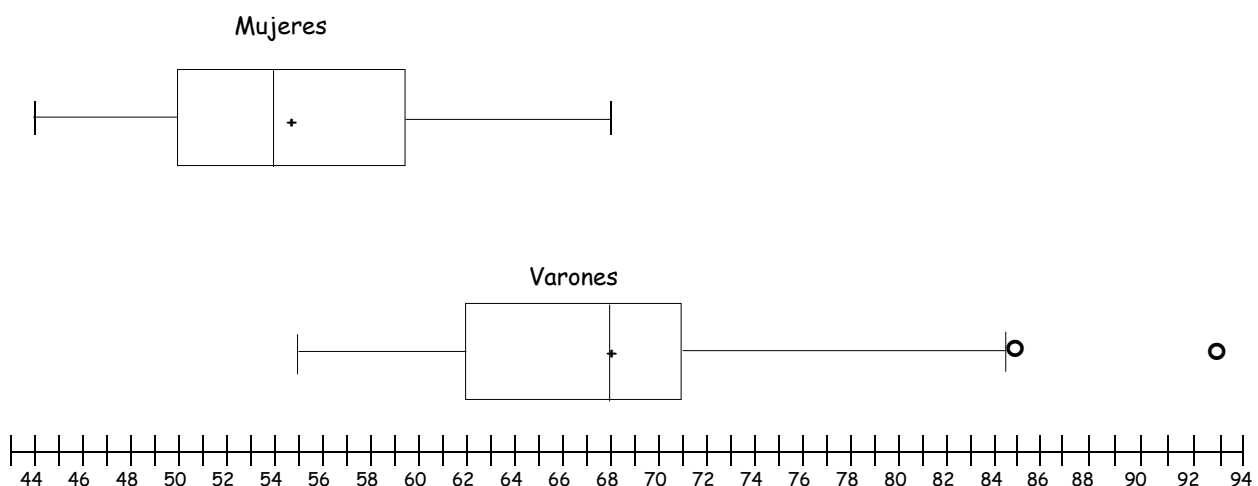
Varones

- En primer lugar, tomaremos la muestra de los varones y anotaremos la información necesaria:
 - Valor mínimo: $x_{\min} = 55$ kg
 - Valor máximo: $x_{\max} = 93$ kg
 - Mediana: $\tilde{x} = 68$ kg
 - Media aritmética: $\bar{x} = 68,083$ 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
- Debemos elegir el valor observado inmediato superior a la REF2 = 48,5 kg para saber hasta dónde llegará la extensión izquierda. En nuestro caso es el valor 55 kg.
- Debemos elegir el valor observado inmediato inferior a la REF3 = 84,5 kg para saber hasta dónde llegará la extensión derecha. En nuestro caso es el valor 75 kg.
- Por lo anterior, vemos que en la muestra de varones, se presentan dos valores atípicos en el extremo superior de la muestra (85 kg y 95 kg).

Mujeres

- Ahora tomaremos la muestra de las mujeres y anotaremos la información necesaria:
 - Valor mínimo: $x_{\min} = 44$ kg
 - Valor máximo: $x_{\max} = 68$ kg
 - Mediana: $\tilde{x} = 54$ kg
 - Media aritmética: $\bar{x} = 54,722$ kg

- o Primer cuartil: $Q_1 = 50$ kg
- o Tercer cuartil: $Q_3 = 60$ kg
- o Rango intercuartílico: $RI = Q_3 - Q_1 = 10$ kg
- o $REF1 = Q_1 - 3.RI = 20$ kg
- o $REF2 = Q_1 - 1.5.RI = 35$ kg
- o $REF3 = Q_3 + 1.5.RI = 75$ kg
- o $REF4 = Q_3 + 3.RI = 90$ kg
- Debemos elegir el valor observado inmediato superior a la $REF2 = 35$ kg para saber hasta dónde llegará la extensión izquierda. En nuestro caso es el valor *44 kg*.
- Debemos elegir el valor observado inmediato inferior a la $REF3 = 75$ kg para saber hasta dónde llegará la extensión derecha. En nuestro caso es el valor *68 kg*.
- Por lo anterior, vemos que en la muestra de mujeres, no se presenta valores atípicos ni anómalos.



1.5 Distribuciones bidimensionales de frecuencias

En algunos estudios estadísticos tomamos, para cada individuo, valores de dos variables estadísticas, por ejemplo, estatura y peso. Este tipo de análisis requiere analizar ambas variables a la vez y su comportamiento en particular. Son múltiples las aplicaciones que requieren de estos estudios, y, a la vez, son muchos los conceptos que se ven involucrados. En este apartado haremos una introducción al análisis de datos en variables bidimensionales.

A. FRECUENCIAS CONJUNTAS, MARGINALES Y CONDICIONADAS

- Si las variables en estudio son X e Y , donde X toma los valores x_1, x_2, \dots, x_n e Y toma los valores y_1, y_2, \dots, y_m . Podemos escribir los datos obtenidos de distintas maneras:
 - En forma de *listado*, que se usa, fundamentalmente, cuando los datos son apareados y la cantidad de valores que toma X es la misma que toma Y .

| X | Y |
|----------|----------|
| x_1 | y_1 |
| x_2 | y_2 |
| \vdots | \vdots |
| x_i | y_i |
| \vdots | \vdots |
| x_n | y_n |

- En forma de *tabla de doble entrada o tabla de contingencia*, que se usa, fundamentalmente, cuando los pares de datos se repiten.

| | y_1 | y_2 | \dots | y_j | \dots | y_m | |
|---------|---------------|---------------|---------|---------------|---------|---------------|--------------|
| x_1 | f_{11} | f_{12} | \dots | f_{1j} | \dots | f_{1m} | $f_{1\cdot}$ |
| x_2 | f_{21} | f_{22} | \dots | f_{2j} | \dots | f_{2m} | $f_{2\cdot}$ |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| x_i | f_{i1} | f_{i2} | \dots | f_{ij} | \dots | f_{im} | $f_{i\cdot}$ |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| \cdot | \cdot | \cdot | \dots | \cdot | \dots | \cdot | \cdot |
| x_n | f_{n1} | f_{n2} | \dots | f_{nj} | \dots | f_{nm} | $f_{n\cdot}$ |
| | $f_{\cdot 1}$ | $f_{\cdot 2}$ | \dots | $f_{\cdot j}$ | \dots | $f_{\cdot m}$ | n |

- Los valores f_{ij} indican las frecuencias absolutas con que aparece el par (x_i, y_j) . Estas frecuencias se llaman *frecuencias conjuntas*.
- Los valores $f_{i\cdot}$ indican la suma de las frecuencias absolutas correspondientes a la i -ésima fila, es decir, las frecuencias correspondientes a $X = x_i$. Es-

tas frecuencias se llaman **frecuencias marginales**. De forma análoga, los valores $f_{.j}$ indican la suma de las frecuencias absolutas correspondientes a la j -ésima columna, es decir, las frecuencias correspondientes a $Y = y_j$.

- Al analizar las frecuencias marginales, surge, naturalmente, la posibilidad de obtener, a partir de la tabla de frecuencias bidimensionales, dos distribuciones unidimensionales, la que corresponde a X y la de Y . A las distribuciones así obtenidas se las llama **distribuciones marginales**.
- El valor n , corresponde a la suma de todas las frecuencias y se llama *gran total*.
- Otro tipo de distribución para la variable X es la que puede obtenerse fijando un valor $Y = y_j$, que se conoce como **distribución de X condicionada para $Y = y_j$** . Análogamente, se puede obtener la **distribución de Y condicionada para $X = x_i$** .

Veremos un ejemplo, donde analizaremos los conceptos expuestos anteriormente:



Ejemplo:

Al clasificar una serie de modelos de automóviles por el número de cilindros y su origen, se obtuvo la siguiente tabla:

Distribución del número de cilindros en una muestra de autos según su origen

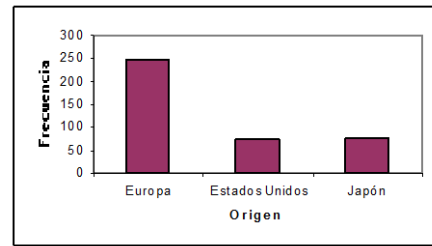
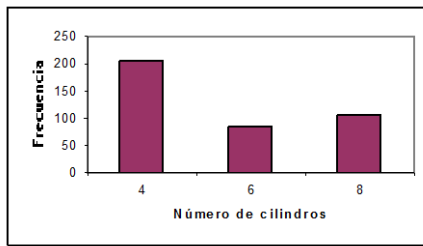
| | | Número de cilindros | | | Total |
|---------------|-----------------------|----------------------------|-----------|------------|--------------|
| | | 4 | 6 | 8 | |
| Origen | Europa | 140 | 57 | 51 | 248 |
| | Estados Unidos | 40 | 12 | 20 | 72 |
| | Japón | 27 | 15 | 36 | 78 |
| | Total | 207 | 84 | 107 | 398 |

A partir de la tabla podemos obtener dos distribuciones marginales, sumando las frecuencias de las filas y las columnas.

Podemos realizar con ellas tablas y gráficas, como vemos a continuación:

| Distribución del número de cilindros según su origen | |
|---|-------------------|
| Número de cilindros | Frecuencia |
| 4 | 207 |
| 6 | 84 |
| 8 | 107 |

| Distribución de los autos | |
|----------------------------------|-------------------|
| Origen | Frecuencia |
| Europa | 248 |
| Estados Unidos | 72 |
| Japón | 78 |

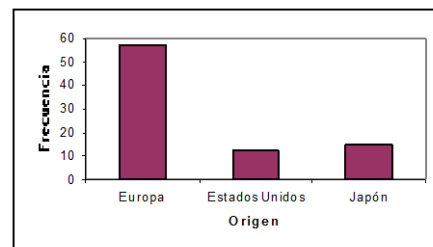
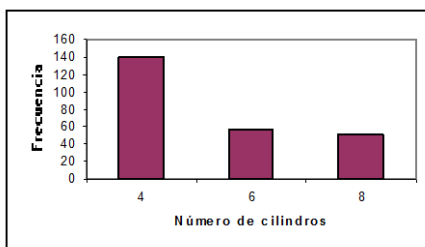


Podemos obtener, también, distribuciones condicionadas, por ejemplo, la distribución del número de cilindros para un origen en especial (lo haremos para Europa) o la distribución del origen para un determinado número de cilindros (lo haremos para 6 cilindros).

Podemos realizar con ellas tablas y gráficas, como vemos a continuación:

| Distribución del número de cilindros para automóviles europeos | |
|--|------------|
| Número de cilindros | Frecuencia |
| 4 | 140 |
| 6 | 57 |
| 8 | 51 |

| Distribución del origen de los autos de seis cilindros | |
|--|------------|
| Origen | Frecuencia |
| Europa | 57 |
| Estados Unidos | 12 |
| Japón | 15 |



Dependencia e Independencia Estadística

Diremos que la variable X es **independiente** de Y si se verifica que:

$$f_{ij} = f_{i.} \cdot f_{.j} / n \quad \text{para todos los pares } (x_i, y_j)$$

Dependencia Funcional y Dependencia Aleatoria

Generalmente, cuando se realiza un estudio estadístico se está interesado en un carácter de los individuos de la población. Una de las preguntas a las cuales se trata de dar respuesta es, si existe alguna relación entre dos variables X y Y .

Para algunos fenómenos, es posible encontrar una fórmula que exprese exactamente los valores de una variable en función de la otra: son los llamados **fenómenos deterministas**. Éste es el caso de **dependencia funcional** entre dos variables. En este tipo de relación, los valores que toma una de las variables que-

dan determinados, de un modo preciso, por los valores que toma la otra variable, que se considera como independiente. Por ejemplo, al estudiar la caída libre de un cuerpo, donde Y representa la distancia recorrida y siendo X el tiempo transcurrido desde su lanzamiento, para una constante $g = 9,8 \text{ m/s}^2$, se sabe que $Y = (1/2).g.X^2$.

Existen muchos fenómenos en los que, al observar pares de valores correspondientes a variables estadísticas, no es posible encontrar una fórmula que relacione, de un modo funcional, esas variables. Si dichos pares de valores son representados en un sistema cartesiano, los puntos, en general, no se ajustan de un modo preciso a una curva plana, sino que se obtiene un conjunto de puntos más o menos dispersos. Una representación de ese tipo recibe el nombre de **nube de puntos, diagrama de dispersión o dispersograma**. Veamos algunos ejemplos:

Figura A

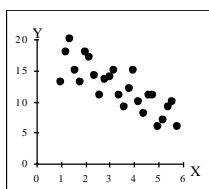


Figura B

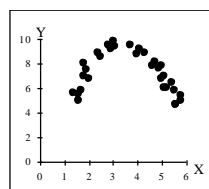
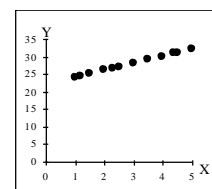


Figura C



Aunque puede apreciarse que en ninguno de los casos es posible encontrar una relación funcional entre las dos variables, sin embargo, observamos una variación conjunta de las variables. En la figura A la relación es inversa, puesto que al crecer X , disminuye Y . En la Figura C, por el contrario, observamos una relación directa, es decir, al crecer X , crece Y . Además, vemos que estos diagramas **'tienen'** a la forma de alguna curva plana, es decir, si bien no es posible dar una fórmula que represente la relación entre las variables, podríamos aproximar la relación entre ellas con una recta de pendiente negativa en la figura A, con una parábola invertida en la figura B y con una recta de pendiente positiva en la figura C. Todos estos ajustes se trabajan con métodos de Cálculo Numérico como aproximación funcional por el método de los mínimos cuadrados.

El Concepto de Asociación

El estudio de la posible relación entre dos variables cuantitativas suele iniciarse mediante la observación del correspondiente diagrama de dispersión. La presencia de una relación entre las variables se pondrá de manifiesto en el diagrama por una cierta tendencia de los puntos a acumularse en las proximidades de una línea, como hemos visto en los ejemplos anteriores.

Al tratar de estudiar si existe o no una relación entre dos variables estadísticas, tratamos de analizar si hay algún tipo de relación entre ellas, si se puede medir la intensidad de esta relación mediante algún coeficiente, si sirve este coeficiente para comparar la intensidad de la relación de diferentes variables, etcétera. Todos estos interrogantes nos llevan a la necesidad de conocer la **asociación** que existe entre las variables estudiadas.

B. CORRELACIÓN Y REGRESIÓN

Hay dos enfoques, íntimamente relacionados en el estudio de las variables aleatorias bidimensionales. El primero se ocupa de dar medidas de la dependencia entre las variables unidimensionales que entran en la variable bidimensional y se denomina **teoría de la correlación**. El segundo trata de dar medios de obtener, aproximadamente, el valor de una de las variables, cuando se da el valor de la otra y se lo llama **teoría de la regresión o ajuste de curvas**.

Más precisamente, los métodos estadísticos presentados hasta ahora se han referido a una sola variable X . Muchos de los problemas del trabajo estadístico, sin embargo, involucran dos o más variables. Veremos el método aplicado al caso de dos variables; pero puede aplicarse a más de dos.

En algunos problemas, las variables se estudian simultáneamente, para ver la forma en que se encuentran interrelacionadas; en otros se tiene una variable de interés particular y las restantes se estudian por la posibilidad de que aclaren aspectos de la primera. Estas dos clases de problemas se conocen, por lo general, con los nombres de **correlación y regresión**, respectivamente.

Correlación Lineal

Un problema de correlación se presenta cuando el individuo se pregunta si existe alguna relación entre un par de variables que le interesan, por ejemplo, ¿existe alguna relación entre el uso del tabaco y las afecciones cardíacas?, ¿entre la capacidad de aprender música y la aptitud científica?, ¿entre la recepción de ondas de radio y la actividad de las manchas solares?, ¿entre la belleza y la inteligencia?



Ejemplo:

Con objeto de ilustrar la forma en que se procede a estudiar la relación entre dos variables, analizaremos un ejemplo sobre los promedios de calificaciones correspondientes al último año de la escuela y al primer año universitario. El promedio de la escuela se designa con X y el promedio universitario con Y .

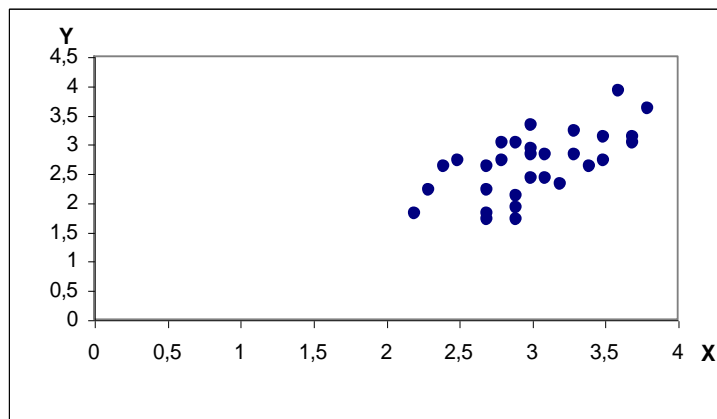
| X | Y |
|-----|-----|
| 3,0 | 2,4 |
| 2,4 | 2,6 |
| 3,7 | 3,0 |
| 3,6 | 3,9 |
| 3,8 | 3,6 |
| 2,9 | 3,0 |
| 3,5 | 3,1 |
| 3,0 | 2,8 |
| 2,3 | 2,2 |
| 3,0 | 2,9 |

| X | Y |
|-----|-----|
| 2,9 | 1,9 |
| 2,7 | 2,2 |
| 3,7 | 3,1 |
| 2,7 | 2,6 |
| 3,3 | 2,8 |
| 2,8 | 2,7 |
| 3,1 | 2,4 |
| 2,8 | 3,0 |
| 3,0 | 3,3 |
| 2,2 | 1,8 |

| X | Y |
|-----|-----|
| 3,1 | 2,8 |
| 3,3 | 3,2 |
| 2,7 | 1,8 |
| 3,5 | 2,7 |
| 2,9 | 2,1 |
| 2,7 | 1,7 |
| 2,9 | 1,7 |
| 3,2 | 2,3 |
| 3,4 | 2,6 |
| 2,5 | 2,7 |

La investigaci n de la relaci n entre las dos variables comienza, generalmente con un intento de descubrir la forma aproximada de la relaci n, marcando los datos, como puntos en el plano cartesiano de coordenadas. Esta gr fica recibe el nombre de *diagrama de dispersi n*. Por este medio, puede decirse con facilidad si existe o no una relaci n acentuada y, en este caso, si puede tratarse como aproximadamente lineal.

El diagrama de dispersi n para las treinta calificaciones se muestra en la figura:



La observaci n de este diagrama de dispersi n muestra que existe una tendencia de los valores bajos de X a estar asociados con los valores bajos de Y, as  como tambi n tienden a asociarse entre s  los valores altos de ambas variables. Por otra parte, y aunque vago, el aspecto general del diagrama de dispersi n es, el de una l nea recta.

Covarianza

En el caso de variables numéricas podemos emplear algunos coeficientes cuyo valor nos indica el tipo de relación entre las variables. El primero de ellos es la **covarianza**, que se indica con S_{xy} y cuya fórmula de cálculo viene dada por la expresión:

$$S_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

La covarianza tiene las siguientes características:

- Es igual a cero si las variables son independientes.
- Es positiva si las variables tienen dependencia directa.
- Es negativa en caso de dependencia inversa.

Coeficiente de Correlación

Un problema de la covarianza es que no hay un máximo para el valor que puede tomar, por lo cual no nos sirve para comparar la mayor o menor intensidad de la relación entre las variables.

Un coeficiente que permite estudiar, no sólo la dirección de la relación, sino también su intensidad, es el **coeficiente de correlación lineal** o *coeficiente de Pearson*, que se define como:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

siendo S_x y S_y , las desviaciones estándar de las variables X e Y en la muestra analizada.

Puesto que las desviaciones estándar son positivas, r tiene el signo de la covarianza y, por tanto:

- Si $r > 0$, la relación entre las variables es directa.
- Si $r < 0$, la relación entre las variables es inversa.
- Si $r = 0$, no existe relación lineal entre las variables.
- Si $r = +1$, la correlación lineal es perfecta y positiva.
- Si $r = -1$, la correlación lineal es perfecta y negativa.

En función de los datos originales, r se define por la siguiente fórmula:

$$r = \frac{n \cdot \sum x_i \cdot y_i - \left(\sum x_i \right) \left(\sum y_i \right)}{\sqrt{n \cdot \sum x_i^2 - \left(\sum x_i \right)^2} \cdot \sqrt{n \cdot \sum y_i^2 - \left(\sum y_i \right)^2}}$$

Si se calcula el coeficiente de correlación a partir de los datos de nuestro ejemplo, se obtendrá $r = 0,63$, este valor de la correlación entre promedios de la escuela y la universidad es típico del coeficiente de correlación entre calificaciones correspondientes a muchos pares de materias. La correlación entre matemática y física puede esperarse un tanto mayor mientras que la correlación entre las calificaciones de matemática y arte dramático se esperaría baja.

Analizaremos dos propiedades de r :

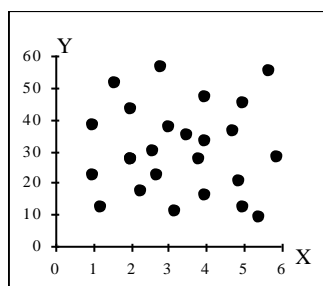
- El valor de r debe satisfacer las desigualdades:

$$-1 \leq r \leq +1$$

- El valor de r será igual a $+1$ o -1 si y sólo si todos los puntos del diagrama se encuentran sobre una línea recta.

Para interpretar el valor de r y descubrir cuáles valores de r son de esperarse en los diversos tipos de relaciones entre X e Y , se presentan algunos diagramas de dispersión con los correspondientes valores de r :

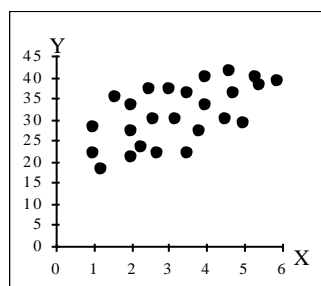
(a)



$r = 0$

(d)

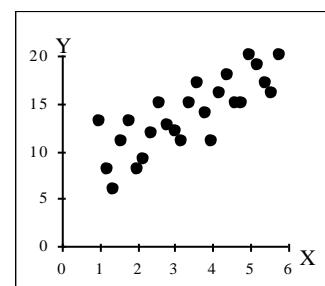
(b)



$r = 0,6$

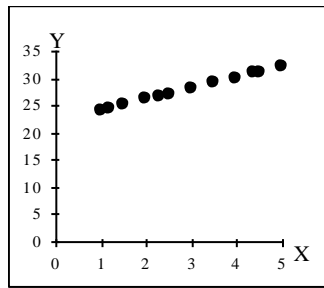
(e)

(c)

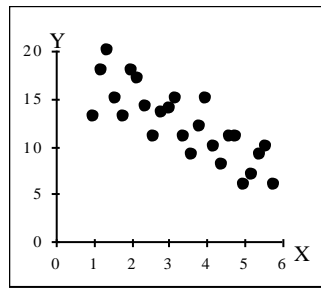


$r = 0,8$

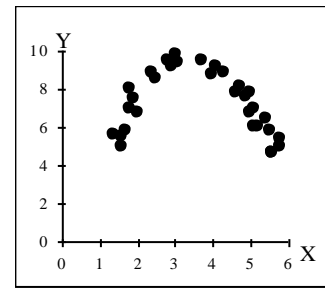
(f)



$r \rightarrow 1$



$r = -0,8$



$r \rightarrow 0$

- o Los diagramas (a), (b), (c) y (d) corresponden a dispersiones con relación lineal cada vez más acentuada.
- o El diagrama (e) es una imagen en espejo del (c), con esto se ve que el valor absoluto de r mide la fuerza de la relación lineal, pero que el signo de r es positivo si Y tiende a crecer al aumentar X , y es negativo si Y tiende a disminuir al crecer X .
- o En el diagrama (f), X e Y están fuertemente relacionadas, pero la relación no es lineal.
- o Este ejemplo indica bien que r es una medida útil para observar lo estrechamente que estén relacionadas dos variables, sólo cuando hay una relación lineal entre ellas.

Interpretación del Coeficiente de Correlación

La interpretación del coeficiente de correlación como medida del grado de relación lineal entre dos variables es una interpretación matemática pura y está completamente desprovista de implicaciones de causa y efecto. El hecho de que dos variables tiendan a aumentar o disminuir al mismo tiempo no implica que una tenga algún efecto directo o indirecto en la otra. Ambas pueden estar sometidas a la influencia de otras variables, de manera que resulten con una estrecha relación matemática, por ejemplo, en un período de varios años el coeficiente de correlación entre los sueldos de maestros y el consumo de licor ha resultado ser de 0,98. Durante este lapso se ha presentado una tendencia ascendente en sueldos y salarios de todos los tipos y una tendencia general a mayores comodidades de vida. En tales condiciones, los salarios de los maestros también habrían de aumentar. Además, la tendencia general de aumento de salarios y poder adquisitivo, así como el aumento de población, se vería reflejada en un aumento en el consumo de licor. Así pues, la alta correlación refleja sólo el efecto común de una tendencia ascendente de las dos variables. Sería incorrecto suponer, por ejemplo, que los maestros gastan su sueldo en la compra de licores.

Los coeficientes de correlación deben manejarse con cuidado si se va a dar una información sensata respecto a la relación entre pares de variables. El utili-

zarlas correctamente requiere familiarización con el campo de aplicación, así como con sus propiedades matemáticas.

Los coeficientes de correlación han probado ser muy útiles para los casos en que es importante determinar la interrelación de algunas variables que se estudian simultáneamente.

- Los valores de r igual a 0,3 y 0,6 sólo significan que tenemos dos correlaciones positivas, una algo más fuerte que la otra.
- Es incorrecto concluir que $r = 0,6$ indica una relación lineal el doble de buena que la que indica el valor $r = 0,3$.
- Al valor de r^2 se lo denomina **coeficiente de determinación** y expresa la proporción de la variación total en los valores de Y que se pueden explicar mediante una relación lineal.
- Una correlación $r = 0,6$ significa que 0,36 ó 36 % de la variación total de los valores de Y en nuestra muestra, se explican mediante una relación lineal con los valores de X .

Regresión Lineal

Es usual estudiar dos o más variables con la esperanza de que cualquier relación que se encuentre pueda usarse para hacer estimaciones o predicciones acerca de una de las variables en particular.

Al estudiar la correlación entre las calificaciones, la intención es, obviamente usar esa relación para tratar de predecir el éxito académico de un estudiante universitario a partir del conocimiento de su promedio en la escuela.

El coeficiente de correlación sólo indica qué tan estrecha es la relación lineal entre las variables y no es capaz de resolver problemas de predicción. De manera semejante, si se calcularan coeficientes de correlación entre promedios de las calificaciones universitarias y resultados de pruebas de aptitud y vocabulario, esas correlaciones sólo servirían para indicar cuáles de estas variables conviene incluir en una función de predicción del éxito académico.

Los métodos que se han desarrollado para manejar problemas de predicción se conocen como métodos de regresión.



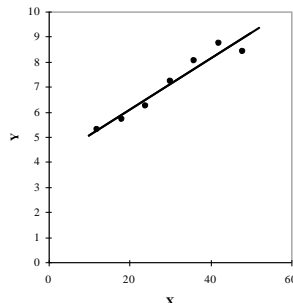
Ejemplo:

Con objeto de explicar los métodos de regresión, consideremos el problema particular de predecir el rendimiento en forraje en función de la cantidad de agua de irrigación aplicada, que se presenta en la siguiente tabla:

| | | | | | | | |
|-----------------|------|------|------|------|------|------|------|
| Agua (X) | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| Rendimiento (Y) | 5,27 | 5,68 | 6,25 | 7,21 | 8,02 | 8,71 | 8,42 |

La cantidad de agua se mide en pulgadas y el rendimiento, en toneladas, de forraje del terreno de una granja experimental.

La gráfica de estos datos es:



Sin duda, surge la necesidad de encontrar la ecuación de esta recta, para calcular, fácilmente, el valor de Y para un valor de X dado.

Veremos un método para encontrar esta ecuación:

La ecuación de la **recta de regresión** es $\hat{y} = a + b \cdot x$

siendo

$$a = \frac{\sum y_i - b \cdot \sum x_i}{n} \quad \text{y} \quad b = \frac{n \cdot \sum x_i \cdot y_i - \left(\sum x_i \right) \left(\sum y_i \right)}{n \cdot \sum x_i^2 - \left(\sum x_i \right)^2}$$

(Éste es el enfoque matricial que da el Cálculo Numérico... ¿Se acuerda?)

En el ejemplo del agua y la ganancia, $a = 3,99$ y $b = 0,103$, luego la recta de regresión de y sobre x que ya fue representada es: $y' = 3,99 + 0,103 \cdot x$

Galton, uno de los pioneros en el campo de la estadística aplicada ha dado este nombre a la recta, en conexión con ciertos estudios que efectuaba para la estimación de la medida en la que *regresa* la estatura del hijo de padre alto, hacia la estatura media de la población, es decir, él descubrió que hijos de padres altos tienen tendencia a ser más bajos que él y los hijos de padres bajos tienen tendencia a ser más altos que su padre.

Interpolación y Extrapolación

Supóngase, que se considera que la relación entre el rendimiento promedio y el agua es estrictamente lineal en esta región de valores de X. Esto significa que, si se hubiera repetido el experimento muchas veces en las mismas condiciones de crecimiento y se hubieran promediado por separado los valores de Y correspondientes a cada uno de los siete valores de X, estos promedios hubiesen dado un

conjunto de puntos casi precisamente sobre una línea recta. Mientras mayor sea el número de estas repeticiones, mayor será la precisión esperada. Esta suposición dice, esencialmente, que existe una línea recta teórica que expresa la relación lineal entre el valor medio teórico de Y y el valor correspondiente de X .

Si se acepta la linealidad, entonces se esperaría que el valor dado por la línea recta ajustada a la muestra, **que es casi igual a 7, fuese más cercano al valor de la línea teórica para $x = 30$ que el valor observado de 7,21**, porque es de esperarse más estabilidad en la línea recta de muestra que en una sola observación puntual. En vista de este razonamiento, se puede predecir el valor de la línea teórica correspondiente a $x = 30$ como el correspondiente valor de y sobre la línea de regresión de muestra.

Si se tiene interés en un valor intermedio de X , que no ha sido analizado en la muestra, se emplea la línea de regresión para encontrar el valor predicho para y , correspondiente al valor de x . Diremos, en este caso, que hacemos interpolación. Sin dudas, la predicción será mejor cuanto más ajustados a la recta se presenten los datos de la muestra. Hacer predicción para valores de X fuera del rango de la muestra se llama extrapolación y, en general, si se supone que la relación es lineal sólo en esta región de valores de X , no es legítimo usar la línea recta para predecir valores de Y fuera de este intervalo de valores de X .

Resumiendo...

En el problema de correlación correspondiente al diagrama de dispersión, los datos consistían en una muestra aleatoria de treinta estudiantes, esto significa que tanto X como Y son variables estadísticas cuyos valores quedan determinados sólo después de obtenida la muestra.

En el problema de regresión, sin embargo, los valores de X se eligieron de antemano, de manera que sólo los valores de Y se han determinado por muestra. Ahora bien, la técnica para ajustar una recta a un conjunto de puntos se puede aplicar independientemente de que los valores de X sean fijados de antemano o se obtengan de muestras al azar. Luego, los métodos de regresión podrían haberse aplicado a datos del tipo considerado en el estudio de correlación. **Por otra parte, la interpretación de r como medida del grado de relación lineal entre dos variables, obviamente, no se aplica, si los valores de X se seleccionan de manera no aleatoria, debido a que el valor de r por lo general dependerá, en buena medida, de la elección de los valores de X .**

Los coeficientes de correlación no se prestan a enunciados cuantitativos, a menos que se asocien con la regresión. Así pues, la correlación por lo general es sólo la primera parte en el estudio de la relación de dos variables, mientras que la regresión es la técnica básica en este tipo de estudios.

1.6 Aspectos éticos

Muy frecuentemente la Estadística es mal utilizada. Los estadísticos deberían tener cuidado para evitar los malos usos. Los docentes deberían advertir a sus alumnos, para evitar mediante la educación, la proliferación de ellos. Los malos usos más frecuentes ocurren por la siguientes razones: datos estadísticos inadecuados, sesgos personales, supuestos falsos, indicación falsa de relaciones, comparaciones impropias y errores en operaciones matemáticas.

El análisis de datos debe basarse en un comportamiento ético intachable porque es muy fácil utilizar las medidas adecuadas para que los datos "digan" lo que el que está realizando el análisis al describir los datos quiere que digan. Además, se presta para realizar gráficos que "dibujen" la realidad deseada, a fin de engañar, a través de la manipulación de los datos.

¡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 gráfico 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?
- ✓ ¿Cómo se puede presentar una distribución bidimensional de variables?
- ✓ ¿Qué es una distribución marginal?
- ✓ ¿Cómo se define la independencia estadística?
- ✓ ¿Es lo mismo dependencia funcional que dependencia estocástica? ¿Por qué?
- ✓ ¿Qué es una distribución condicionada?
- ✓ Defina el concepto de asociación.
- ✓ ¿Cuándo y para qué se utiliza un análisis de regresión?
- ✓ ¿Cuándo y para qué se utiliza un análisis de correlación?
- ✓ ¿Qué diferencia hay entre regresión y correlación?
- ✓ ¿Qué es la covarianza?
- ✓ ¿Cuáles son las características del coeficiente de correlación?
- ✓ ¿Cómo se debe interpretar el coeficiente de correlación?
- ✓ ¿Qué es interpolar? ¿Qué es extrapolar?
- ✓ ¿Son ambas acciones válidas para un análisis estadístico?



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...



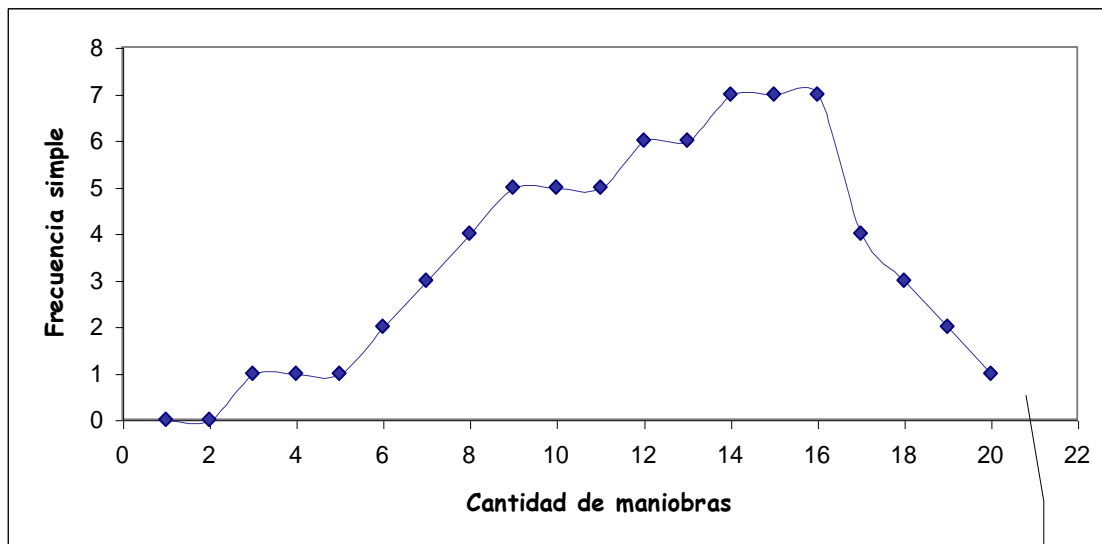


Respuestas

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"?

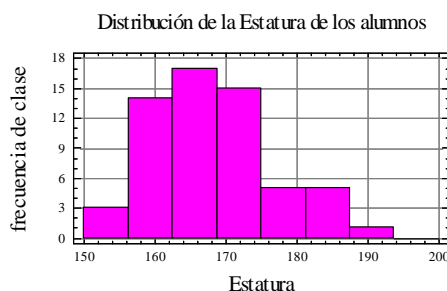
La respuesta depende de cuán machista o feminista sea el que conteste, pero tratando de encontrar un punto de equilibrio, y en base al mito popular, podemos decir que una gráfica más o menos representativa sería:



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 gráficos, 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 de Polimodal.



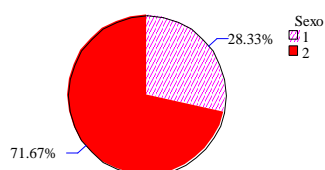
- ¿Qué tipos de chistes causarían más efecto, los machistas o los feministas?

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

Distribución de frecuencias del SEXO de los alumnos

| Sexo | Valor | Frecuencia | | Acumulada | |
|--------|-------|------------|----------|-----------|----------|
| | | 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:



- Respecto a la tabla DEPORTES:
¿Cómo definimos la variable que se refiere a la práctica deportiva?
Cantidad de veces que realiza actividades deportivas.

Tabla de frecuencia para la variable DEPORTE

| Deporte | Valor | Frecuencia | | Acumulativa | |
|------------------|-------|------------|----------|-------------|----------|
| | | Absoluta | Relativa | Absoluta | 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

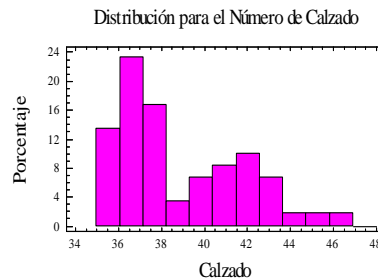
SISTEMÁTICAMENTE: Dos o más veces por semana

¿Cuál es la escala de medición?

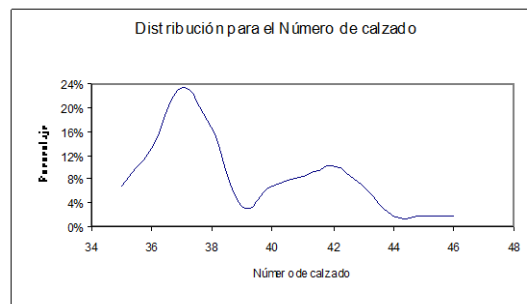
La variable es cualitativa, medida en escala ordinal.

- ¿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 sexo, 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.

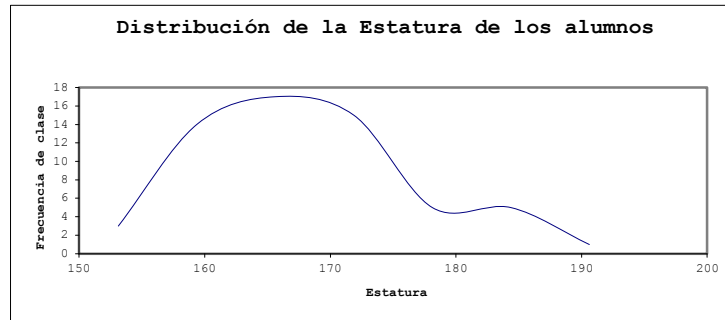
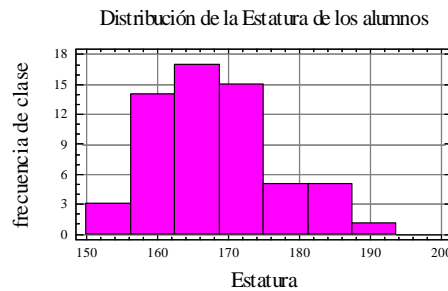
- ¿Es coherente la distribución del número de calzados con el sexo de los estudiantes? ¿Por qué?

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 medio.

¿Qué medida de tendencia central daría usted si...

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

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

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.*

A trabajar solos...

La precipitación anual de lluvias, aproximada a décima de centímetro, para un período de 30 años es 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 ||

$$x_{\min} = 28,3$$

$$x_{\max} = 47,5$$

$$R = x_{\max} - x_{\min} = 19,2$$

$$k = 1 + 3,3 \cdot \log n \approx 5,8745 \approx 6$$

$$l = R / k = 19,2 / 6 = 3,2$$

| 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:

$$\bar{x} = 37,58 \text{ cm}$$

La precipitación anual de lluvias promedio es de 37,58 cm.

Mediana:

$$^{\circ} Me = (n+1) / 2 = 31 / 2 = 15,5 \Rightarrow Me \in [34,7 ; 41,1)$$

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

Esto indica que el 50% de los años la precipitación anual fue de 36,792 cm o menos y en el otro 50% la precipitación anual fue de 36,792 cm o más.

Modo:

$$Mo \in [34,7 ; 37,9)$$

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

Siendo:

$$\Delta_1 = 9 - 4 = 5$$

$$\Delta_2 = 9 - 1 = 8$$

La precipitación anual más frecuente es de 35,93 cm.

Primer cuartil:

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

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

Esto indica que 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:

$$^{\circ}D_4 = 4 \cdot (n+1) / 10 = 12,4 \Rightarrow D_4 \in [34,7 ; 41,1)$$

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

Esto indica que 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:

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

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

Esto indica que 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:

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

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.

