Estadística Descriptiva

M. Carmen Carollo Limeres

Estadística descriptiva univariable

Algunos conceptos básicos

DATOS

Resultados de una medición (peso de los conejos de una camada)

Resultados de un proceso de conteo (nº de eucaliptos en Galicia)

Los datos contienen información:

- **¿Cómo se obtienen?**
- **¿Cómo se analizan?**
- **¿Cómo se interpretan?**

ESTADÍSTICA: Disciplina que se ocupa de:

- La recolección, organización, resumen y análisis de datos ***(Est. descriptiva*)** y la
- La obtención de conclusiones para la población a partir de la muestra ***(Est. inferencial*)**

BIOESTADÍSTICA:
Cuando los datos que se analizan proceden de las ciencias biológicas o médicas.

La **estadística descriptiva** se ocupa de recoger, clasificar y resumir la información contenida en los datos a través de técnicas numéricas y gráficas, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen.

POBLACIÓN: Es el colectivo de individuos al cual se refiere el estudio que se pretende realizar.

- **Ejemplo:** El población de vacas de la comunidad gallega en 2003.

MUESTRA: Es una parte representativa de la población a estudiar.

Ejemplo: Un investigador está interesado en el estudio de un tipo de miel que se produce en la provincia de Lugo. En 1990, se recogieron 66 muestras de miel en dos zonas de dicha provincia clasificándolas de acuerdo a las siguientes variables: *Zona, Humedad, Cenizas, Solins (sólidos insolubles), Azured (azucares redutores), Sacar (sacarosa), Conduc (conductividad), AcTot (acidez total), AcLib (acidez libre), AcLab (acidez lactónica), HMF* (*Hidroximetilfusfural,* importante para clasificar las mieles), *pH, Lac_Lib (AcLac/AcLib)* e *Azu_Hum (Azured/Humedad).*

Estadística. FBA I. Curso 2011-2012

	zona	humedad	cenizas	solins	azured	sacar	conduc	actot	aclib	aclac	hmf	ph	laclib	azuhum
1	1	16,6	,73	,046	71,9	2,09	183	30,3	29,44	,83	7,87	4,53	,03	4,33
2	1	16,3	,43	,009	70,9	,75	251	43,5	41,80	1734,00	9,70	4,19	,04	4,35
3	1	17,6	,83	,034	70,8	,61	167	35,2	30,03	5,12	10,56	4,26	,17	4,02
4	1	16,8	,58	,013	71,1	1,24	198	32,8	32,13	,71	17,28	4,20	,02	4,23
5	1	16,5	,44	,010	68,1	2,78	234	34,9	31,82	3,03	16,32	4,20	,10	4,13
6	1	17,0	,61	,023	70,7	3,01	154	41,7	34,77	6,94	4,22	3,96	,20	4,16
7	1	16,2	,26	,011	70,6	1,48	239	39,6	38,97	,60	10,94	3,79	,02	4,36
8	1	17,3	,41	,020	66,6	5,10	361	28,4	27,11	1,31	6,72	4,56	,05	3,85
9	1	18,8	,60	,010	67,2	1,33	222	28,3	27,46	,80	11,50	4,41	,03	3,68
10	1	17,1	,39	,005	66,8	1,23	417	27,6	27,03	,59	15,00	4,77	,02	3,91
11	1	17,3	,43	,060	73,3	1,50	310	43,4	36,95	6,44	10,50	4,26	,17	4,24
12	1	17,3	,23	,045	70,1	,98	272	34,9	30,52	4,42	11,70	4,10	,14	4,05
13	1	16,4	,48	,015	70,9	2,93	276	37,8	36,56	1,27	7,10	4,55	,03	4,32
14	1	16,6	,24	,005	69,1	1,63	460	41,6	38,39	3,20	11,90	4,36	,08	4,16
15	1	17,6	,28	,010	68,3	,20	326	43,3	36,64	6,66	10,18	4,01	,18	3,88
16	1	18,2	,69	,035	69,6	,50	178	39,3	31,82	7,52	16,32	4,15	,24	3,82
17	1	17,8	,41	,076	69,8	,70	185	21,4	16,94	4,43	4,42	4,22	,26	3,92
18	1	17,8	,17	,066	67,4	,65	378	33,1	27,03	6,02	7,30	4,09	,22	3,78
19	1	16,6	,78	,126	67,2	,40	133	39,3	34,09	5,85	11,52	3,97	,17	4,05
20	1	16,2	,37	,020	71,4	1,85	239	32,4	27,42	4,94	11,52	4,13	,18	4,41
21	1	16,6	,31	,011	68,8	1,98	340	41,4	38,28	3,11	8,45	4,30	,08	4,14
22	1	15,7	,33	,050	68,7	1,30	242	32,1	29,24	2,90	2,88	4,19	,01	4,37
23	1	17,2	,20	,017	69,9	,98	644	26,7	23,07	3,61	5,57	4,57	,16	4,06
24	1	17,8	,27	,027	68,2	2,58	288	46,7	42,07	4,60	10,40	3,87	,11	3,83
25	1	17,4	,03	,017	71,8	2,50	110	35,7	32,88	2,77	5,20	3,59	,08	4,12
26	1	16,3	,05	,083	65,6	,59	251	26,6	24,87	1,73	4,80	3,52	,07	4,02
27	1	19,4	,10	,142	66,5	1,25	550	37,2	31,15	6,04	8,45	3,87	,19	3,43
28	1	16,9	1,01	,036	61,7	2,50	66	24,6	21,67	2,96	5,57	3,55	,14	3,64
29	1	18,6	,45	,011	68,8	2,14	316	22,8	19,00	3,84	4,03	4,52	,20	3,70
30	1	17,4	,31	,022	67,3	1,09	302	31,5	29,79	1,67	8,83	4,35	,06	3,86

VARIABLE: Característica que queremos estudiar (sexo, nivel de estudios, nº hermanos, peso...)

<table>
<tr><td rowspan="4">Variables</td><td rowspan="2">Cualitativas</td><td>Nominales</td></tr>
<tr><td>Ordinales</td></tr>
<tr><td rowspan="2">Cuantitativas</td><td>Discretas</td></tr>
<tr><td>Continuas</td></tr>
</table>

Ejemplo

En la última hora han acudido al servicio de urgencias de un hospital ocho pacientes, cuyos datos de ingreso se encuentran resumidos en la siguiente tabla.

Sexo	Peso	Estatura	Temperatura	Nº de visitas	Dolor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

Variables: Sexo, peso, estatura(m),temperatura, nº de visitas previas al servicio de urgencias y dolor)

DISTRIBUCIÓN DE FRECUENCIAS

Frecuencia Absoluta: es el número de veces que ocurre cada resultado (x_i). La denotamos por n_i

Frecuencia Relativa: Es la frecuencia absoluta dividida por el número de observaciones. La denotamos por $f_i = n_i / N$

Frecuencia Absoluta Acumulada: Es el número de veces que se ha observado el resultado x_i o valores anteriores. La denotamos por $N_i = n_1 + n_2 + ... + n_i$.

Frecuencia Relativa Acumulada: Es la frecuencia absoluta acumulada dividida por el número total de observaciones. La denotamos por $F_i = N_i / N$.

Las frecuencias verifican las siguientes **propiedades:**

- Frecuencias absolutas: $0 \leq n_i \leq N$ $\quad \sum n_i = N$
- Frecuencias relativas $0 \leq f_i \leq 1$ $\quad \sum f_i = 1$
- Frecuencias absolutas acumuladas:

$$0 \leq N_i \leq N \qquad N_k = N$$

- Frecuencias relativas acumuladas:

$$0 \leq F_i \leq 1 \qquad F_k = 1$$

Tipo de variable		
	Cualitativas nominales	Frecuencia absoluta
		Frecuencia relativa
		Porcentaje
	Cualitativas ordinales y Cuantitativas	Frecuencias acumuladas
		Frec relativas acumuladas
		Porcentaje acumulado

Tabla de frecuencias: Se utiliza para representar la información de una muestra de tamaño N extraída de la población

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
⋮	⋮	⋮	⋮	⋮
x_k	n_k	N_k	f_k	F_k
	$\sum n_i = N$		$\sum f_i = 1$	

Tabla de Frecuencias para una Variable Cualitativa

Modalidades	Frecuencias Absolutas	Frecuencias Relativas
A_1	n_1	f_1
A_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
A_k	n_k	f_k
	$\sum n_i = N$	$\sum f_i = 1$

Tabla de Frecuencias para una Variable Cuantitativa discreta

Valores	Frecuencias Absolutas	Frec. Absol. Acumuladas	Frecuencias Relativas	Frec.Relat. Acumuladas
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
⋮	⋮	⋮	⋮	⋮
x_k	n_k	N_k	f_k	F_k
	$\sum n_i = N$	$N_k = N$	$\sum f_i = 1$	$F_k = 1$

Tabla de Frecuencias para una Variable Cuantitativa Continua

Agrupar los valores que puede tomar la variable en intervalos y contar el número de veces que la variable cae en cada intervalo.

Intervalo de clase.

Marca de clase.

Número de intervalos:

- Cuantos menos intervalos, menos información se recoge.
- Cuantos más intervalos, más difícil es manejar las frecuencias.
- El entero más próximo a $\sqrt{N}$ (10 como máximo)

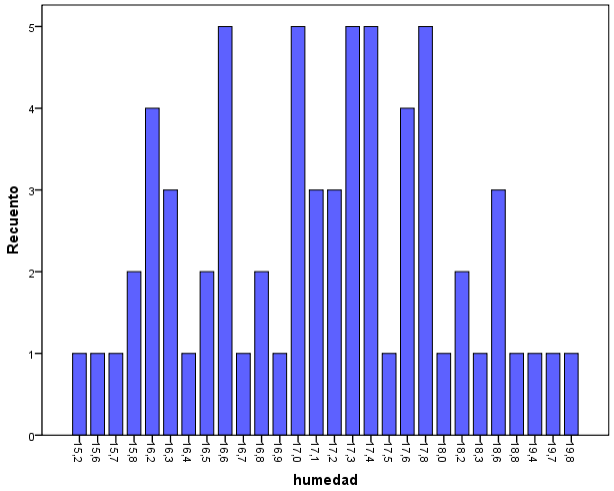
A**mplitud (tamaño o longitud)** del intervalo.

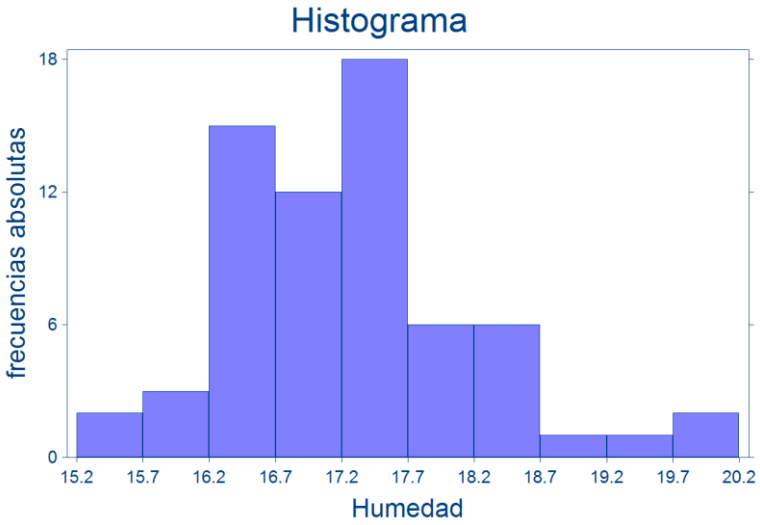
Ejemplo: Con objeto de hacer un estudio sobre las alturas de los alumnos de uno de los grupos de prácticas, se recogen los datos de los alumnos del grupo G1**.**

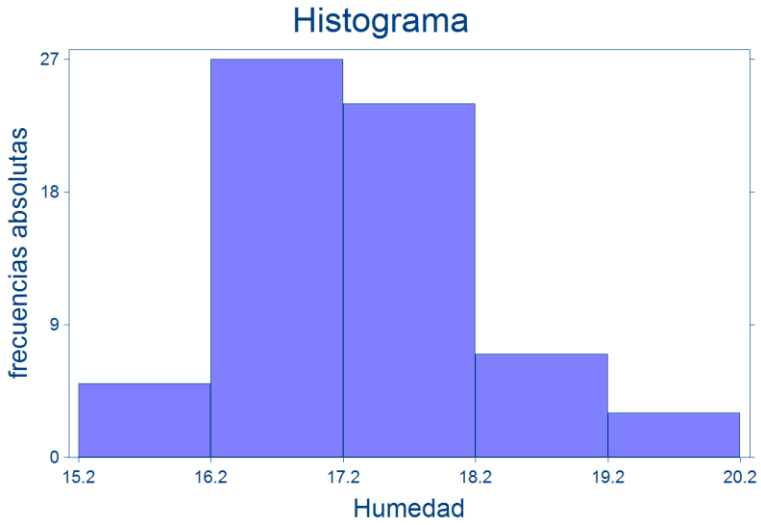
INTERVALOS	MARCAS DE CLASE	FREC. ABS.	FREC. ABS. ACUMULADA	FREC REL.	FREC. REL. ACUMULADA
$[L_0,L_1)$	X_1	n_1	N_1	f_1	F_1
$[L_1,L_2)$	X_2	n_2	N_2	f_2	F_2
⋮	⋮	⋮	⋮	⋮	⋮
$[L_{k-1},L_k)$	x_k	n_k	N_k	f_k	F_k
		$\sum n_i = N$		$\sum f_i = 1$	

Obtención de Intervalos

1. Rango: $RV = V_{MAX} - V_{MIN}$
2. Número de intervalos a considerar: $\sqrt{N}$
3. Amplitud: $\dfrac{RV}{\text{Nº intervalos}}$
4. Frecuencias = Número de observaciones por intervalo


Recuento
humedad
0
1
2
3
4
5
15.2
15.6
15.7
15.8
16.2
16.3
16.4
16.5
16.6
16.7
16.8
16.9
17.0
17.1
17.2
17.3
17.4
17.5
17.6
17.8
18.0
18.2
18.3
18.6
18.8
19.4
19.7
19.8


Histograma
frecuencias absolutas
Humedad
0
6
12
18
15.2
15.7
16.2
16.7
17.2
17.7
18.2
18.7
19.2
19.7
20.2


Histograma
frecuencias absolutas
Humedad
0
9
18
27
15.2
16.2
17.2
18.2
19.2
20.2

Representaciones Gráficas de las Distribuciones de una variable

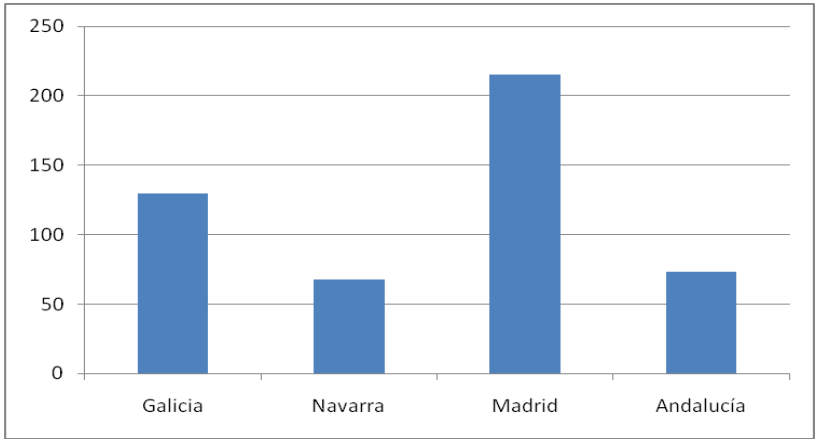
Tipo de variable			
	Cualitativa		Gráfico de rectángulos
			Gráfico de sectores
			Perfil ortogonal
			Pictogramas
	Cuantitativa	*Discreta*	Diagrama de barras o de frec.
			Polígono de frecuencias
			Diag. Frec. acumuladas
		Continua	Histograma
			Polígono frecuencias
			Políg frec acumuladas

Representaciones Gráficas para Variables Cualitativas

1. Diagrama de Rectángulos

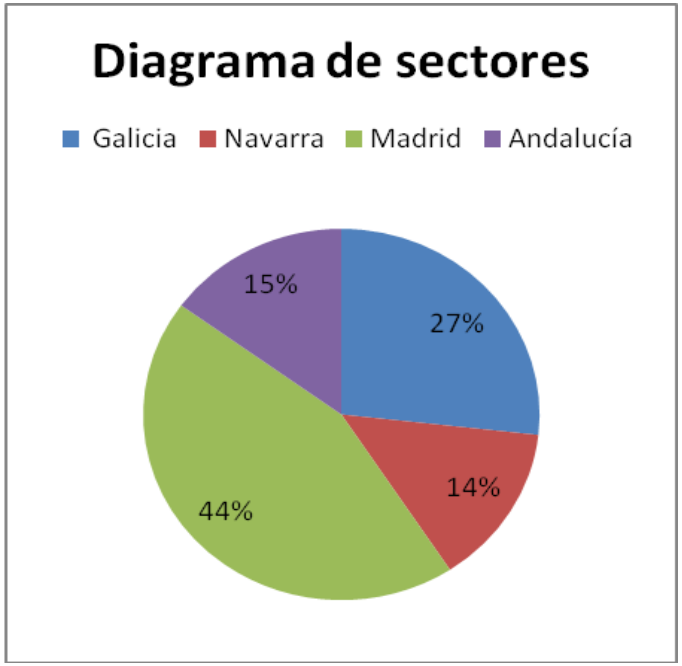
Ejemplo: Para el control de las vacas locas en el año 2003 se tomó una muestra en cuatro comunidades autónomas obteniéndose los siguientes datos.

Comunidad	Nº de animales
Galicia	130
Navarra	68
Madrid	215
Andalucía	73



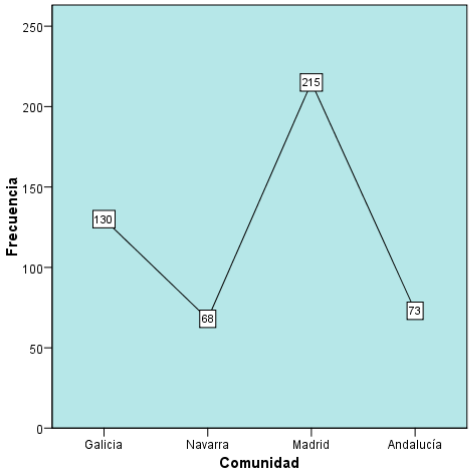
2. **Diagrama de Sectores**

Comunidad	Nº de animales
Galicia	130
Navarra	68
Madrid	215
Andalucía	73



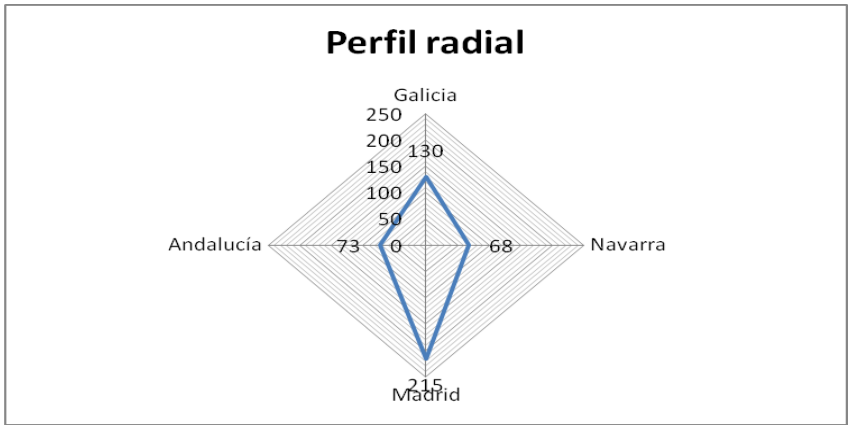
3. **Perfil Ortogonal**

Comunidad	Nº de animales
Galicia	130
Navarra	68
Madrid	215
Andalucía	73



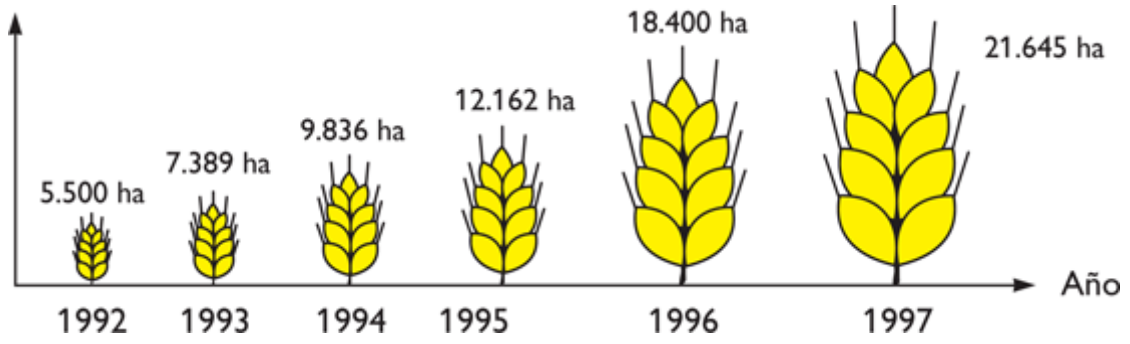
4. Perfil Radial

Comunidad	Nº de animales
Galicia	130
Navarra	68
Madrid	215
Andalucía	73



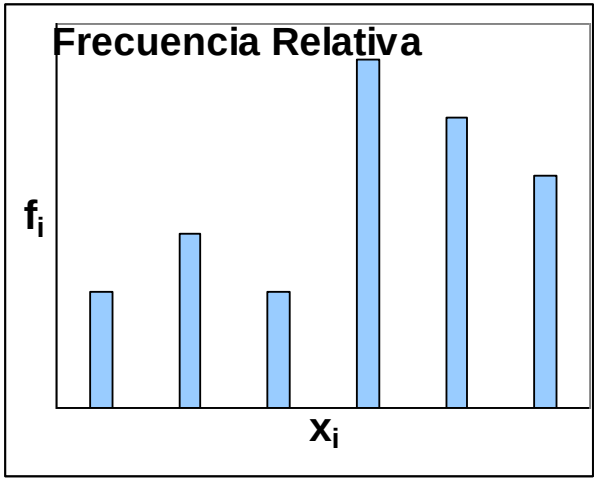
5. Pictogramas

Ejemplo: Plantaciones de trigo

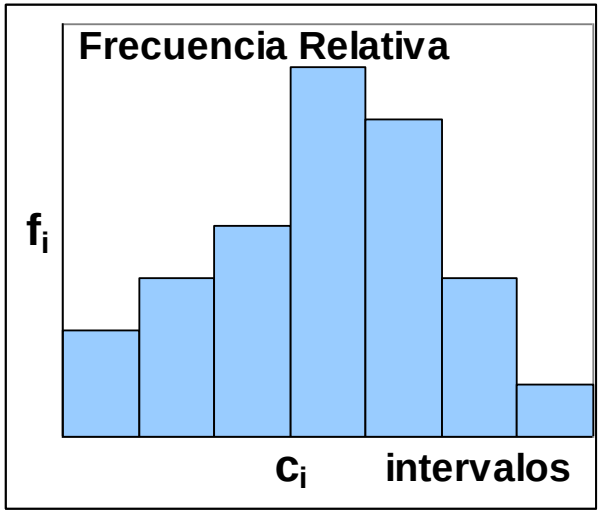


Representaciones Gráficas para Variables Cuantitativas

1. Diagrama de Barras

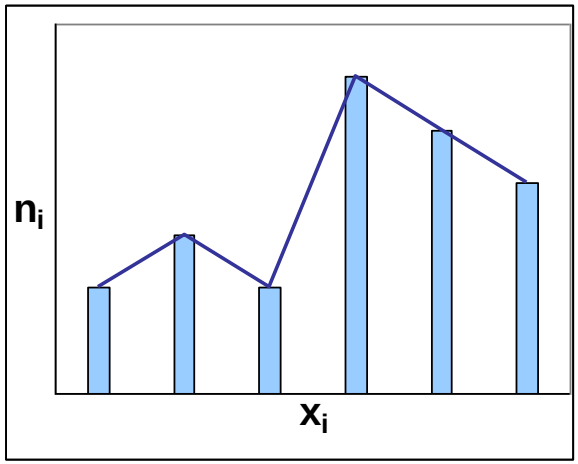


2. Histograma

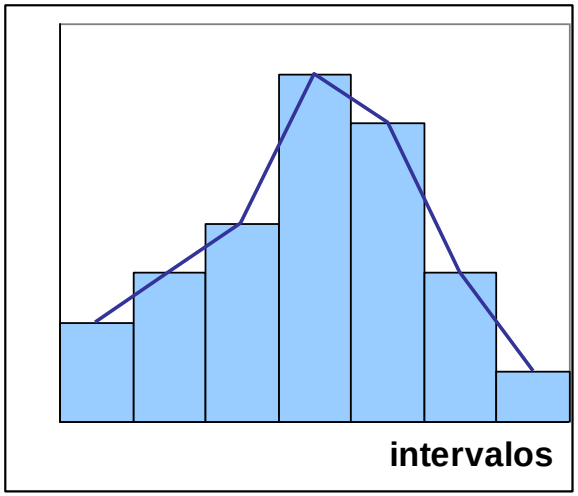


2. Polígono de Frecuencias

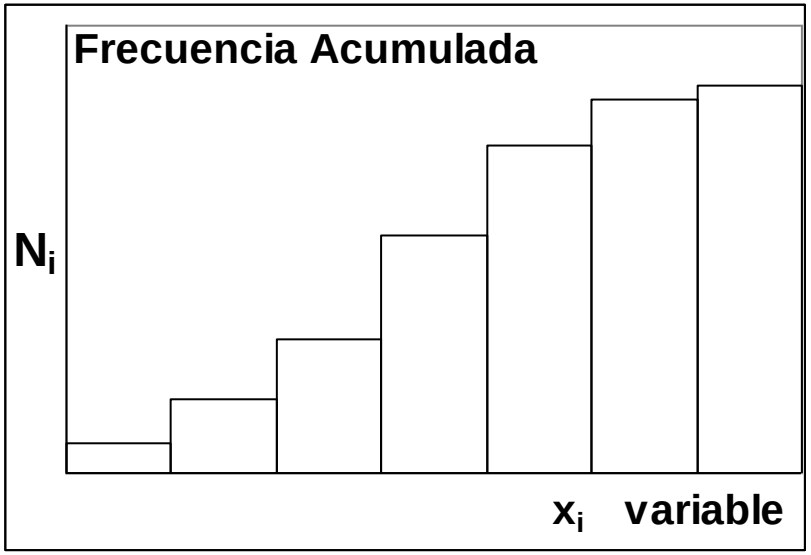
Variable discreta



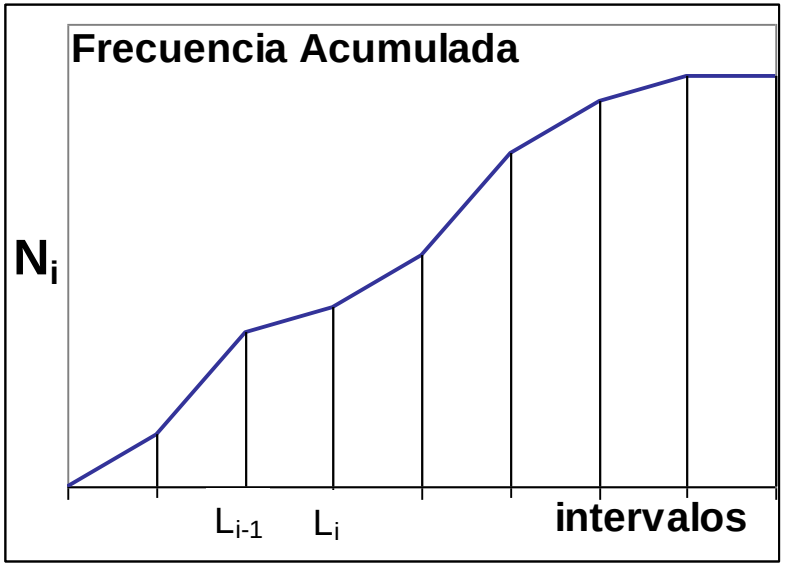
Variable continua



3. Diagrama de frec. Acumuladas
(Diagrama de Barras Acumulado)



4. Polígono de frec. Acumuladas



Ejercicio:

Para evaluar la viabilidad de un proyecto de reforestación de una zona sometida a una fuerte actividad turística, se analiza la composición en mg por cm^3 de desechos orgánicos del territorio. Los datos que se obtienen son los siguientes:

10.8 9.1 22.5 12.3 17.3 31.5 17.9 16.7 20.3
19.2 23.8 25.5 15.4 20.3 2.3 13.5 9.3 2.72
10.9 25.9

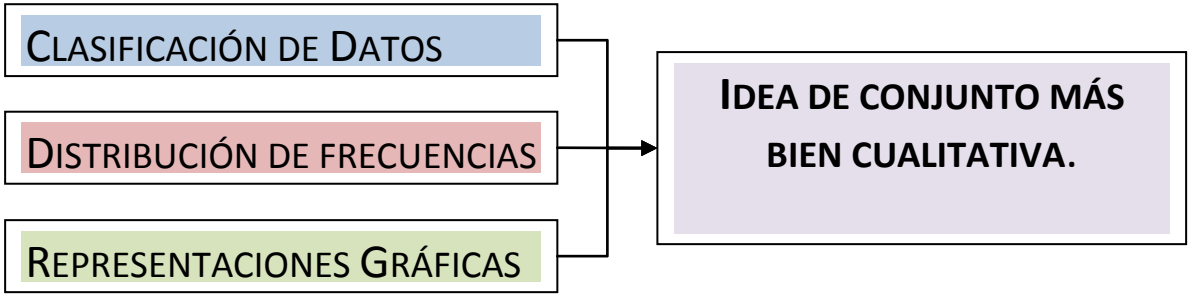
Obtener la correspondiente tabla de frecuencias y representar dichas frecuencias mediante un histograma y un polígono de frecuencias acumuladas.

6. Diagrama de Tallo y Hojas: Para **variables cuantitativas discretas y continuas**. Se redondean los datos a 2 ó 3 cifras significativas, se toman como tallos la primera o las dos primeras cifras y como hojas las últimas cifras. Se representan, separados por una línea vertical, los tallos a la izquierda y las hojas a la derecha del tallo correspondiente. Así cada tallo se representa una única vez y define una clase. El número de hojas de cada tallo representa su frecuencia.

Ejemplo: A partir de las edades observadas de un grupo de individuos: 38, 49, 49, 44, 48, 49, 48, 59, 53, 51, 52, 53, 54, 56, 56, 58, 57, 52, 53, 50, 61, 69, 62, 68, 62, 63, 64, 69, 53, 50, 61, 69, 62, 68, 62, 63, 64, 69, 68, 63, 66, 66, 68, 68, 68, 66, 76, 74, 70, 72, 70, 76, 74, 74,77, 71, 77, 79, 70, 76, 80, 83

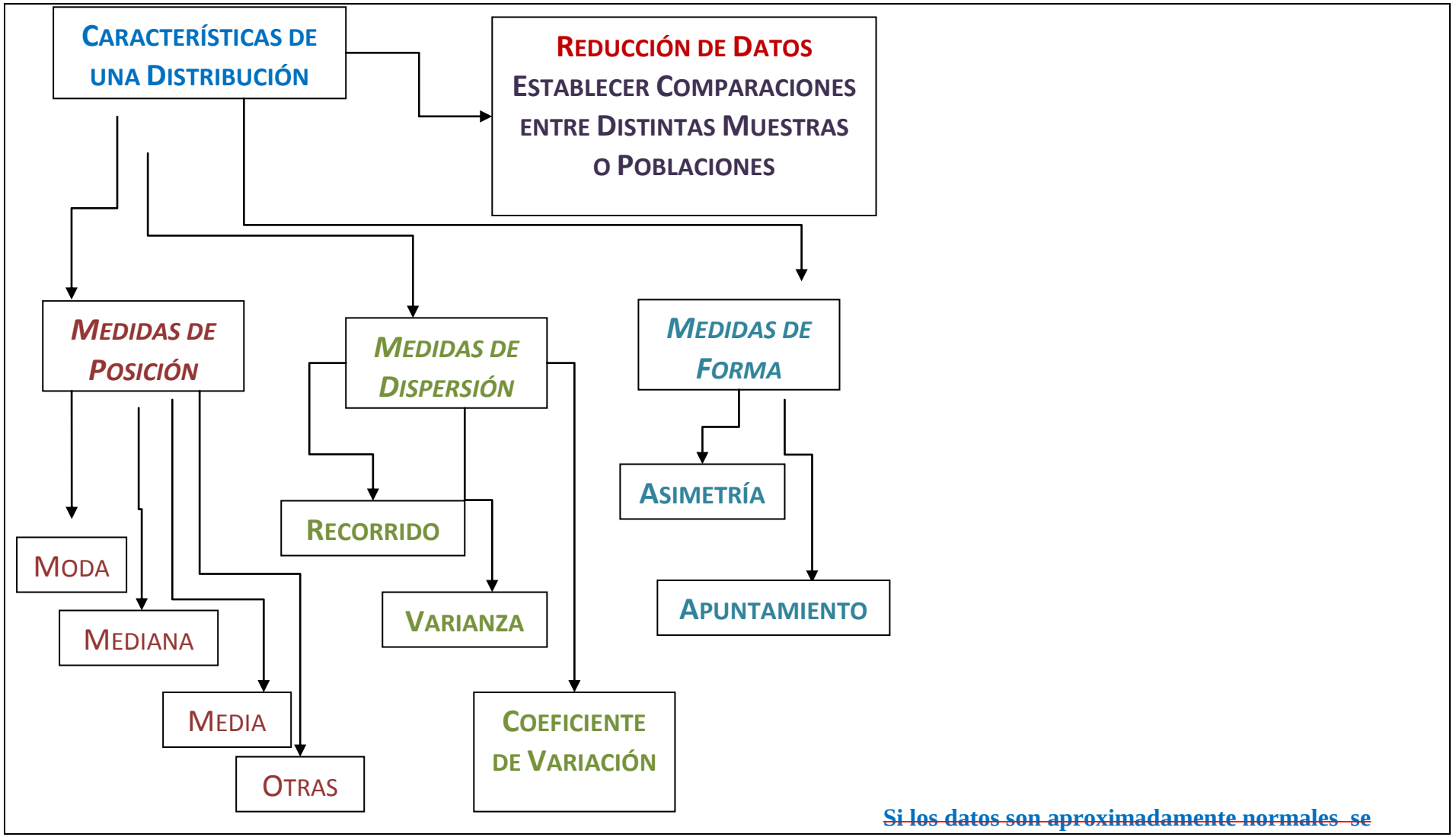
1	3	8
6	4	488999
15	5	001223333466789
21	6	112222334466688888999
14	7	00012444666779
2	8	03

Medidas Representativas de una Distribución de Frecuencias



Resumen mediante el uso de valores numéricos que den idea de:

- La ubicación o el centro de los datos **(Medidas de localización o posición).**
- La concentración de los datos alrededor de dicho centro **(Medidas de dispersión).**
- Otros rasgos de la distribución **(Medidas de forma).**



Si los datos son aproximadamente normales se

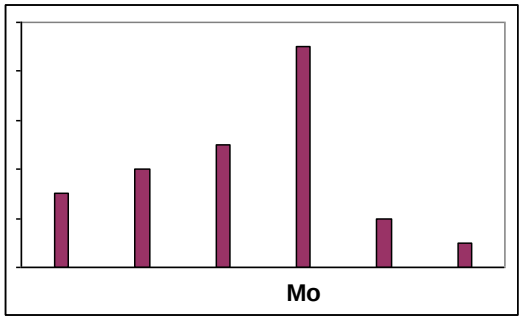
Medidas de Posición

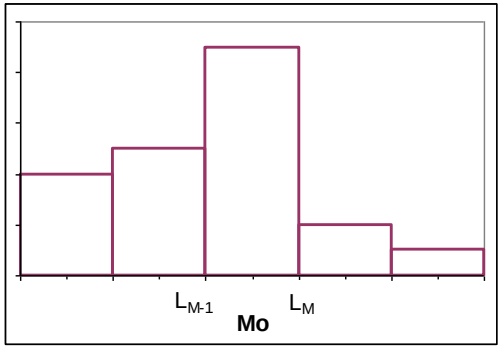
- **tendencia central:** moda, mediana y media
- **tendencia no central**: cuantiles (cuartiles, deciles, percentiles)

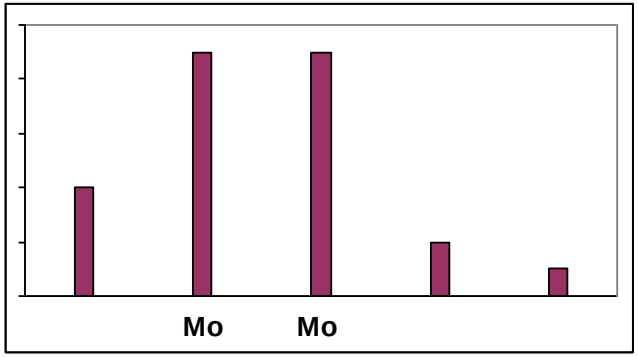
MODA: Es el valor/es o modalidad/es que se presenta con mayor frecuencia: **Mo**

Ejemplo: Con el fin de controlar la contaminación de un río, todas las semanas se hace una medición del nivel de ácido úrico, las mediciones durante cinco semanas fueron: **10, 14, 14, 12, 13**

Mo=14







MEDIANA: Valor de la variable ordenada que deja a su izquierda el mismo nº de observaciones que a su derecha (Me)

Ejemplo: 10, 14, 14, 12, 13

Ordenamos los datos:

10, 12, 13, 14, 14

Me=13

Ejemplo: 10, 11, 12, 13, 14, 14

Me=(12+13)/2 = 12.5

MEDIA ARITMÉTICA:

$$\bar{X} = \frac{\text{suma de observaciones}}{\text{número de observaciones}} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

Ejemplo: 10, 14, 14, 12, 13

$$\bar{X} = \frac{10 + 2*14 + 12 + 13}{5} = 12,6$$

Propiedades de la media:

- Tiene las mismas unidades que los datos.
- Su valor está comprendido entre el mínimo y el máximo de los datos.
- Si se multiplican todas las observaciones por una cantidad, a, y se les suma otra cantidad, b, la media es igual a: $a\bar{X}$ + b.
- Es el centro de gravedad de los datos: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
- Está afectada por cada valor. Valores extremos pueden distorsionarla.

Propiedades de la mediana:

- Tiene las mismas unidades que los datos. Su valor está comprendido entre el mínimo y el máximo de los datos.
- Los valores extremos no tienen efectos importantes sobre ella. En distribuciones simétricas se debe utilizar la media, mientras que en distribuciones asimétricas debemos utilizar la mediana.

Ejemplo: Con objeto de hacer un estudio sobre las alturas de los alumnos de uno de los grupos de prácticas, se recogen los datos de los alumnos del grupo G1.

INTERV.	x_i	n_i	N_i
[160,170)	165	5	5
[170,180)	175	6	11
[180,190)	185	3	14
[190,200)	195	2	16

MODA: intervalo de mayor frecuencia: [170,180) $\text{Mo}=\frac{170+180}{2}=175$

MEDIA: $\bar{X}=\frac{165\cdot 5+175\cdot 6+185\cdot 3+195\cdot 2}{16}=176.25$

Medidas de posición de tendencia no central

CUANTILES:

- **CUARTILES** : Dividen la muestra en cuatro partes iguales
- **DECILES :** Dividen la muestra en diez partes iguales
- **PERCENTILES:** Dividen la muestra en cien partes iguales

Ejemplo: 10, 10, 11, 12, 12, 12, 13, 14, 14, 15, 15

Primer cuartil: Q1=11	**Segundo decil : D2=10**
Segundo cuartil= Mediana: Q2=12	**Cuarto decil: D4=12**
Tercer cuartil: Q3=14	**Noveno decil: D9=15**

Medidas de dispersión

x_i	n_i
0	1
500	1
1000	1

$$\bar{X} = \frac{0+500+1000}{3} = 500$$

ALTA VARIABILIDAD

x_i	n_i
501	1
499	1

$$\bar{X} = \frac{499+501}{2} = 500$$

BAJA VARIABILIDAD

a) Medidas de dispersión absolutas:

- RECORRIDO o RANGO: $$R = V_{MAX} - V_{MIN}$$

-VARIANZA:

$$s^2 = \frac{\sum_{i=1}^{k}\left(x_i - \bar{X}\right)^2 n_i}{N} = \sum_{i=1}^{k}\left(x_i - \bar{X}\right)^2 f_i = \sum_{i=1}^{k} x_i^2 f_i - \bar{X}^2$$

- $s^2 \geq 0$

-DESVIACIÓN TÍPICA:

$$s = +\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k}\left(x_i - \bar{X}\right)^2 n_i}{N}}$$

-RANGO INTERCUARTÍLICO (RIC): RIC= Q3 –Q1

b) Medidas de dispersión relativa: Para comparar la dispersión de variables que aparecen en unidades distintas o que toman valores de magnitudes muy diferentes, utilizaremos el:

-COEFICIENTE DE VARIACIÓN: $\mathrm{CV} = \frac{S}{\bar{X}}$

b) Otras medidas de dispersión

-DESVIACIÓN ABSOLUTA RESPECTO A LA MEDIA

$$D_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- DESVIACIÓN ABSOLUTA RESPECTO A LA MEDIANA $D_{Me} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me|$

-MEDIANA DE LA DESVIACIÓN ABSOLUTA $MEDA = Mediana\left\{|x_i - Me|, i = 1,\ldots,n\right\}$

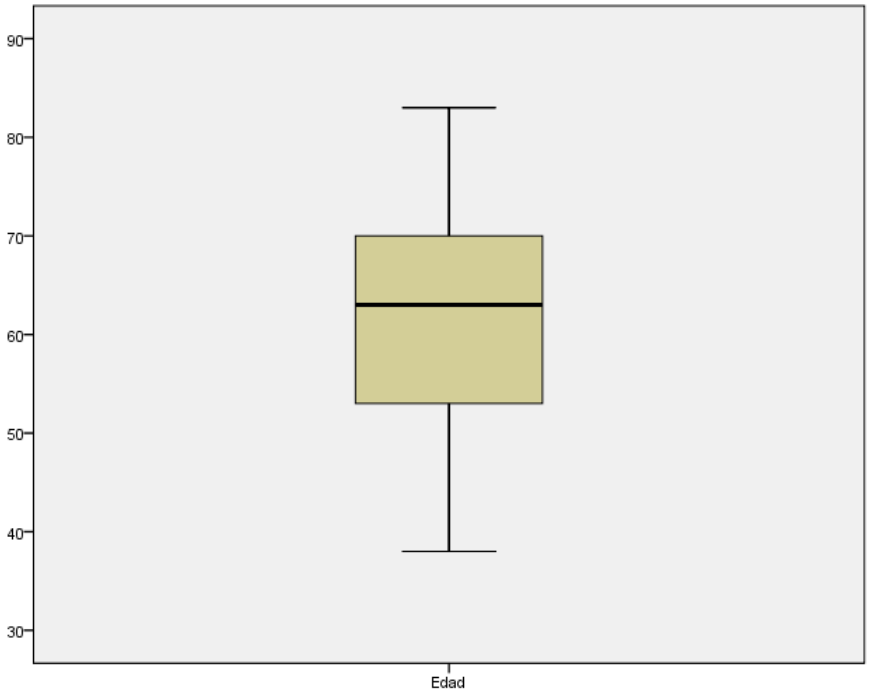
DIAGRAMA DE CAJA: Es una presentación visual que describe la dispersión y simetría. El diagrama consta de una caja delimitada por los cuartiles Q1 y Q3 y en cuyo interior se representa la mediana. De los extremos salen unas líneas (bigotes) que se extienden hasta:

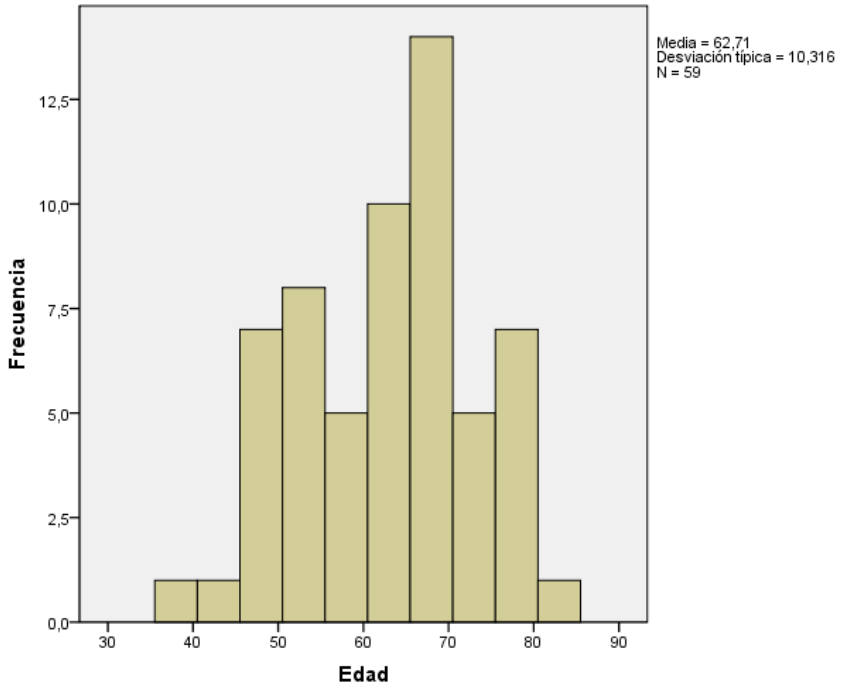
LI=máx { menor observación , Q1-1.5*RIC }

LS=mín { mayor observación , Q3+1.5*RIC }

Dato atípico: Cualquier dato que no se encuentre dentro de este rango

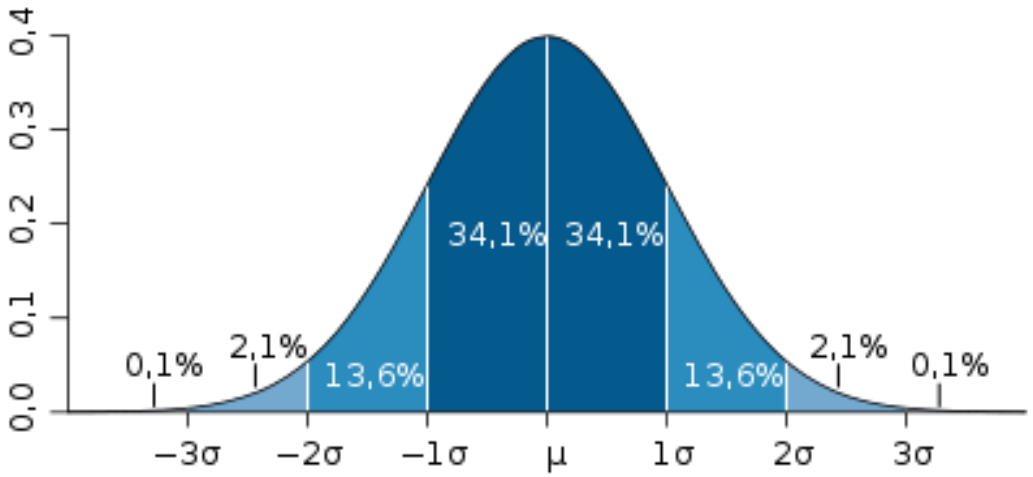
1	3. 8
6	4. 488999
15	5. 001223333466789
21	6. 112222334466688888999
14	7. 00012444666779
2	8. 03





Si los datos son aproximadamente normales se suelen resumir con la media y la varianza.
Si no lo son se resumen a través de 5 números: valor mínimo, primer cuartil, mediana, tercer cuartil y valor máximo.

Si se trata de normales estándar, el 95% de las observaciones están entre -2 y +2.



Ejemplo: **Un laboratorio nacional y otro extranjero producen un insecticida para eliminar la plaga del pulgón en las plantas. El nacional garantiza una efectividad de 5 horas, desviación típica de 1,5 hora, mientras que el internacional nos da una efectividad media de 6 y una desviación típica de 2,4 hora. ¿Qué laboratorio tiene menor variación? Si después de su aplicación se ha comprobado que la efectividad del primero es de 5,5 horas y la del segundo es de 6,4 ¿cuál tiene más efectividad?**

Coeficiente de variación G1 = 0,3 **Coeficiente de variación G2 = 0,4** $\mathrm{CV} = \frac{S}{\overline{X}}$

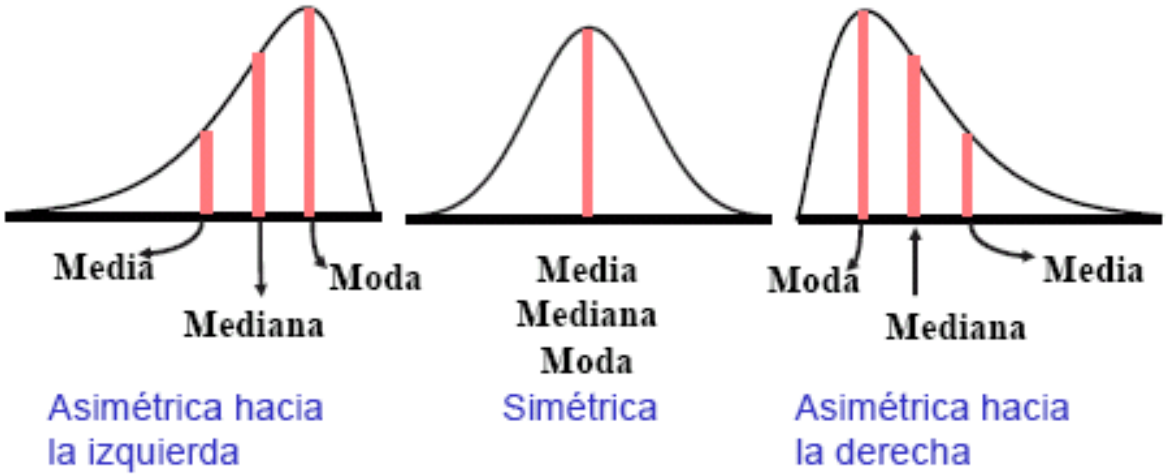
Para comparar individuos de distintos grupos es necesario la tipificación de los datos. Para ello se les resta la media de su grupo y se divide por la correspondiente desviación típica. $X_{tip} = \frac{X - \overline{X}}{S}$

Valores tipificados: **(5,5 – 5)/1,5 =0,33** **(6,4-6)/2,4 = 0,16**

Medidas de forma

-COEFICIENTE DE ASIMETRÍA:

$$CA = \sum_{i=1}^{n} \left(x_i - \bar{X} \right)^3 \frac{1}{ns^3}$$



-COEFICIENTE DE CURTOSIS:

Mide el grado de concentración, respecto a un estándar, de los datos en la región central de la distribución. Este estándar es la 'normal'.