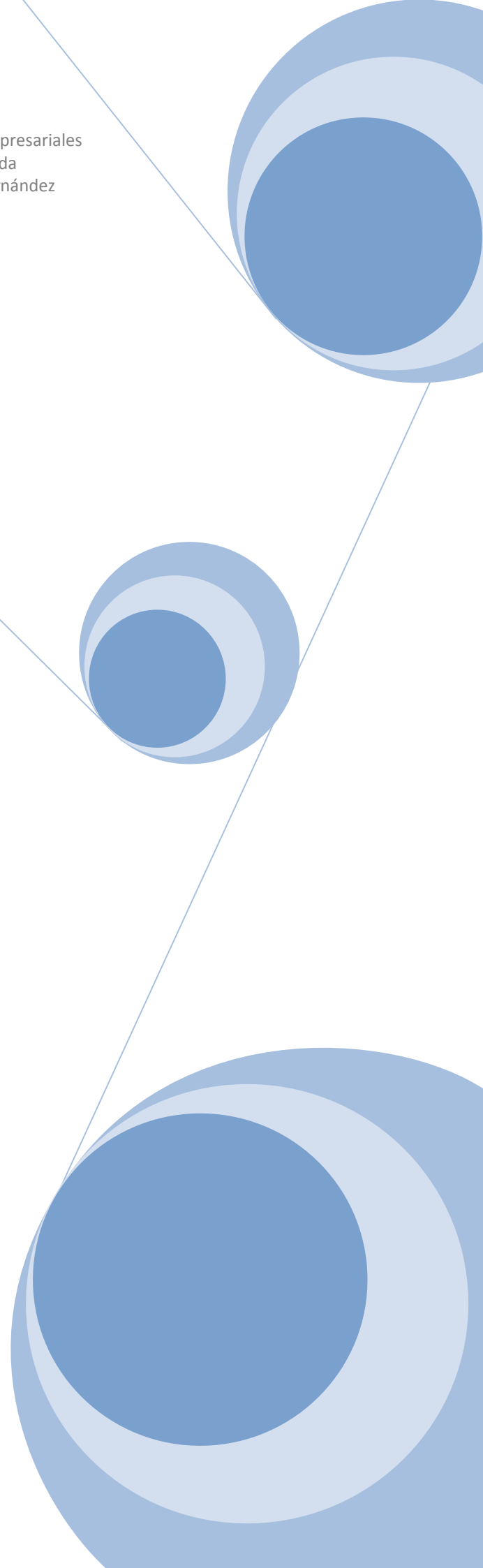




Grado Administración y Gestión
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

ESTADÍSTICA UNIDIMENSIONAL



TABLAS DE FRECUENCIAS - MEDIDAS DE POSICIÓN

x_i	n_i	N_i	$f_i = n_i / N$	$F_i = N_i / N$
1	2	2	0,05	0,05
2	6	8	0,15	0,2
3	10	18	0,25	0,45
4	5	23	0,125	0,575
5	10	33	0,25	0,825
6	3	36	0,075	0,9
7	2	38	0,05	0,95
8	2	40	0,05	1
$N = \sum_{i=1}^8 n_i = 40$			$\sum_{i=1}^8 f_i = 1$	

Frecuencia absoluta (n_i): Número de veces que se repite el dato x_i

Frecuencia absoluta acumulada: N_i

Frecuencia relativa: $f_i = n_i / N$

Frecuencia relativa acumulada: $F_i = N_i / N$

Mediana: Se divide el número de datos entre dos ($N/2$), se va a la columna N_i (frecuencia absoluta acumulada), si el dato se encuentra allí, la MEDIANA es el valor de la x_i correspondiente. Si el valor ($N/2$) no se encuentra en la columna N_i , la MEDIANA es el valor de la x_i correspondiente a la N_i inmediata superior (ordenada). Se denota por M_e

Mediana: $M_e = 4$

Cuartiles (Q_k): El Cuartil k -ésimo se calcula de forma similar a la mediana M_e , aunque el cálculo se hace sobre $(k \cdot N/4)$, donde $k = 1, 2, 3, 4$

Tercer Cuartil: $Q_3 = 5$

Percentiles (P_k): El Percentil k -ésimo se calcula de forma similar a la mediana M_e , aunque el cálculo se hace sobre $(k \cdot N/100)$, donde $k = 1, 2, \dots, 100$

Percentil 95: $P_{95} = 7$

Recorrido: La diferencia entre el valor máximo y el valor mínimo de una variable.

$R = \max(x_i) - \min(x_i) = 8 - 1 = 7$

Recorrido Intercuartílico : Diferencia entre el tercer cuartil (Q_3) y el primer cuartil (Q_1)

$$R_I = Q_3 - Q_1 = 5 - 3 = 2$$

Recorrido Semiintercuartílico : La mitad del recorrido intercuartílico. $R_{SI} = \frac{Q_3 - Q_1}{2}$

$$R_{SI} = \frac{Q_3 - Q_1}{2} = \frac{5 - 3}{2} = 1$$

REPRESENTACIONES GRÁFICAS

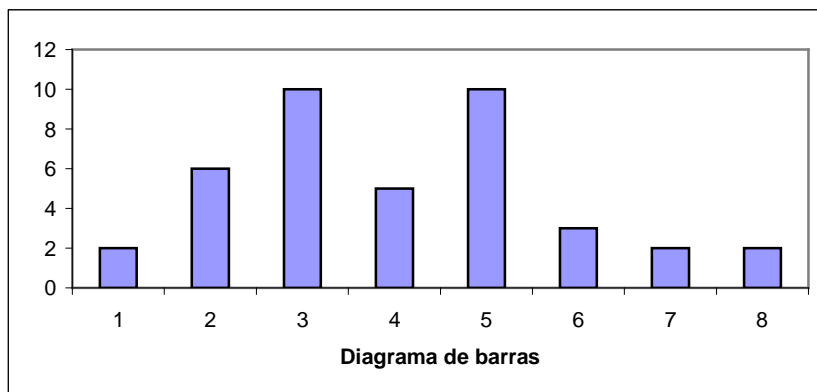
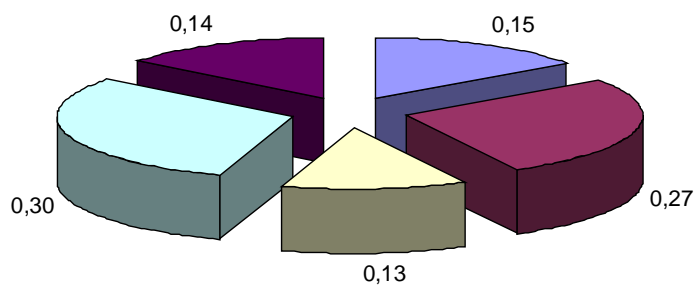


DIAGRAMA DE SECTORES

Carreras	Alumnos	Fr.relative	Grados Sector
Matemáticas	2136	0,15	54,06
Filosofía	3870	0,27	97,95
Derecho	1830	0,13	46,32
Económicas	4328	0,30	109,54
Químicas	2060	0,14	52,14
	14224	1	360



MEDIDAS DE TENDENCIA CENTRAL

A veces conviene reducir la información obtenida a un solo valor o a un número pequeño de valores para facilitar la comparación entre distintas muestras o poblaciones. Estos *valores que de alguna forma centralizan la información, reciben el nombre de medidas de tendencia central.*

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$\frac{n_i}{x_i}$	$\log x_i$	$n_i \cdot \log x_i$
1	2	2	2	2	0	0
2	6	12	24	3	0,301	1,806
3	10	30	90	3,333	0,477	4,771
4	5	20	80	1,250	0,602	3,010
5	10	50	250	2	0,699	6,990
6	3	18	108	0,500	0,778	2,334
7	2	14	98	0,286	0,845	1,690
8	2	16	128	0,250	0,903	1,806
	40	162	780	12,619	4,606	22,408

$$\text{Media Aritmética: } \bar{x} = \frac{\sum_{i=1}^8 x_i \cdot n_i}{N} = \frac{162}{40} = 4,05$$

$$\text{Media Armónica: } \bar{x}_A = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_8}{x_8}} = \frac{N}{\sum_{i=1}^8 \frac{n_i}{x_i}} = \frac{40}{12,619} = 3,17$$

La relación entre las medias es: $\bar{x}_A \leq \bar{x}_G \leq \bar{x} \leftrightarrow 3,17 \leq 3,63 \leq 4,05$

Cuando el valor de $x_i = 0$ la media armónica no tiene sentido.

Media Geométrica:

$$\bar{x}_G = \sqrt[40]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_8^{n_8}} \mapsto \log \bar{x}_G = \frac{1}{40} \sum_{i=1}^8 (n_i \cdot \log x_i) = \frac{1}{40} (22,408) = 0,5602$$

$$\log \bar{x}_G = 0,5602 \mapsto \bar{x}_G = 10^{0,5602} = 3,63$$

PROPIEDADES DE LA MEDIA ARITMÉTICA

CAMBIO DE ORIGEN

x_i	n_i	$x_i \cdot n_i$
1	2	2
2	6	12
3	10	30
4	5	20
5	10	50
6	3	18
7	2	14
8	2	16
	40	162

$x_i + 3$	n_i	$(x_i + 3) \cdot n_i$
4	2	8
5	6	30
6	10	60
7	5	35
8	10	80
9	3	27
10	2	20
11	2	22
	40	282

CAMBIO DE ESCALA

$4 \cdot x_i$	n_i	$4 \cdot x_i \cdot n_i$
4	2	8
8	6	48
12	10	120
16	5	80
20	10	200
24	3	72
28	2	56
32	2	64
	40	648

$$\bar{x} = \frac{\sum_{i=1}^8 x_i \cdot n_i}{N} = \frac{162}{40} = 4,05$$

$$\frac{\sum_{i=1}^8 (x_i + 3) \cdot n_i}{N} = \frac{282}{40} = 7,05 = (4,05 + 3) = \bar{x} + 3$$

Si se produce un cambio de origen (b), la media aritmética (\bar{x}) se ve afectada en la medida en que se produce el cambio, es decir, ($\bar{x} + b$)

$$\frac{\sum_{i=1}^k (x_i + b) \cdot n_i}{N} = \frac{\sum_{i=1}^k x_i \cdot n_i}{N} + \frac{\sum_{i=1}^k b \cdot n_i}{N} = \frac{\sum_{i=1}^k x_i \cdot n_i}{N} + \frac{b \cdot \sum_{i=1}^k n_i}{N} = \frac{\sum_{i=1}^k x_i \cdot n_i}{N} + b = \bar{x} + b$$

$$\frac{\sum_{i=1}^8 (4 \cdot x_i) \cdot n_i}{40} = \frac{648}{40} = 16,2 = 4 \cdot (4,05) = 4 \bar{x}$$

Si se produce un cambio de escala (c), la media aritmética (\bar{x}) se ve afectada en la medida en que se produce el cambio, es decir, ($c \cdot \bar{x}$)

$$\frac{\sum_{i=1}^k (c \cdot x_i) \cdot n_i}{N} = \frac{c \cdot \sum_{i=1}^k x_i \cdot n_i}{N} = c \cdot \bar{x}$$

CAMBIO DE ORIGEN Y DE ESCALA

x_i	n_i	$x_i \cdot n_i$
1	2	2
2	6	12
3	10	30
4	5	20
5	10	50
6	3	18
7	2	14
8	2	16
	40	162

$4 \cdot x_i + 3$	n_i	$(4 \cdot x_i + 3) \cdot n_i$
7	2	14
11	6	66
15	10	150
19	5	95
23	10	230
27	3	81
31	2	62
35	2	70
	40	768

$$\frac{\sum_{i=1}^8 (4 \cdot x_i + 3) \cdot n_i}{N} = \frac{768}{40} = 19,2 = 4 \cdot (4,05) + 3 = 4 \cdot \bar{x} + 3$$

Si se produce simultáneamente un cambio de origen (b) y de escala (c) en los datos, estos afectan de igual medida a la media aritmética, es decir, $(c \cdot \bar{x} + b)$

$$\frac{\sum_{i=1}^k (c \cdot x_i + b) \cdot n_i}{N} = \frac{\sum_{i=1}^k c \cdot x_i \cdot n_i}{N} + \frac{\sum_{i=1}^k b \cdot n_i}{N} = \frac{c \cdot \sum_{i=1}^k x_i \cdot n_i}{N} + \frac{b \cdot \sum_{i=1}^k n_i}{N} = c \cdot \bar{x} + b$$

MOMENTOS

Un momento de orden r respecto al parámetro c, se define:

$$M_r(c) = \frac{\sum_{i=1}^k (x_i - c)^r \cdot n_i}{N}$$

En particular, interesan dos casos:

- Momentos respecto al origen ($c = 0$): $a_r = \frac{\sum_{i=1}^k (x_i - 0)^r \cdot n_i}{N} = \frac{\sum_{i=1}^k x_i^r \cdot n_i}{N}$
- Momentos respecto a la media ($c = \bar{x}$): $m_r = \frac{\sum_{i=1}^k (x_i - \bar{x})^r \cdot n_i}{N}$

MOMENTOS RESPECTO AL ORIGEN

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
1	2	2	2
2	6	12	24
3	10	30	90
4	5	20	80
5	10	50	250
6	3	18	108
7	2	14	98
8	2	16	128
	40	162	780

$$a_0 = \frac{\sum_{i=1}^8 x_i^0 \cdot n_i}{N} = \frac{\sum_{i=1}^8 n_i}{N} = \frac{40}{40} = 1$$

$$a_1 = \bar{x} = \frac{\sum_{i=1}^8 x_i \cdot n_i}{N} = \frac{162}{40} = 4,05$$

$$a_2 = \frac{\sum_{i=1}^8 x_i^2 \cdot n_i}{N} = \frac{780}{40} = 19,5$$

MOMENTOS RESPECTO A LA MEDIA ($\bar{x} = 4,05$)

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$
1	2	2	2	-3,05	9,303	-6,1	18,605
2	6	12	24	-2,05	4,203	-12,3	25,215
3	10	30	90	-1,05	1,103	-10,5	11,025
4	5	20	80	-0,05	0,002	-0,25	0,012
5	10	50	250	0,95	0,903	9,5	9,025
6	3	18	108	1,95	3,803	5,85	11,408
7	2	14	98	2,95	8,703	5,9	17,405
8	2	16	128	3,95	15,603	7,9	31,205
	40	162	780			0	123,9

$$m_0 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^0 \cdot n_i}{N} = \frac{\sum_{i=1}^8 n_i}{N} = \frac{40}{40} = 1$$

$$m_1 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^1 \cdot n_i}{N} = \frac{0}{40} = 0$$

$$m_2 = s^2 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{123,9}{40} = 3,0975 \text{ (varianza)}$$

La varianza también se puede expresar: $m_2 = s^2 = a_2 - (a_1)^2 = a_2 - (\bar{x})^2$

en efecto, $m_2 = s^2 = 3,0975 = 19,5 - (4,05)^2$

$$m_2 = s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{\sum_{i=1}^k (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \cdot n_i}{N} = \frac{\sum_{i=1}^k x_i^2 \cdot n_i - 2\bar{x} \sum_{i=1}^k x_i \cdot n_i + \bar{x}^2 \sum_{i=1}^k n_i}{N} = \frac{\sum_{i=1}^k x_i^2 \cdot n_i}{N} - 2\bar{x} \cdot \frac{\sum_{i=1}^k x_i \cdot n_i}{N} + \bar{x}^2 \cdot \frac{\sum_{i=1}^k n_i}{N} = a_2 - 2 \cdot \bar{x} \cdot \bar{x} + \bar{x}^2 = a_2 - \bar{x}^2 = a_2 - a_1^2$$

MEDIDAS DE DISPERSIÓN O CONCENTRACIÓN

Las medidas de tendencia central reducen la información de una muestra a un solo valor, pero, en algunos casos, éste valor estará más próximo a la realidad de las observaciones que en otros. Las medidas de dispersión o concentración se encargan de cuantificar la representatividad de estos valores centrales. Resaltar que los términos concentración y dispersión pueden ser utilizados indistintamente, se observa la relación:

alta dispersión ↔ baja concentración baja dispersión ↔ alta concentración

x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$ x_i - \bar{x} $	$ x_i - \bar{x} \cdot n_i$
1	2	2	2	3,05	6,1
2	6	12	24	2,05	12,3
3	10	30	90	1,05	10,5
4	5	20	80	0,05	0,25
5	10	50	250	0,95	9,5
6	3	18	108	1,95	5,85
7	2	14	98	2,95	5,9
8	2	16	128	3,95	7,9
	40	162	780		58,3

$$a_1 = \bar{x} = \frac{\sum_{i=1}^8 x_i \cdot n_i}{N} = \frac{162}{40} = 4,05$$

$$a_2 = \frac{\sum_{i=1}^8 x_i^2 \cdot n_i}{N} = \frac{780}{40} = 19,5$$

$$\sigma^2 = a_2 - (a_1)^2 = 19,5 - (4,05)^2 = 3,0975$$

$$\sigma = \sqrt{3,0975} = 1,76$$

La varianza σ^2 también se define: $m_2 = s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{N}$, al ser suma de cuadrados tomará

siempre valores positivos. En el caso que $s^2 = 0$ se entiende que todos los x_i coinciden con la media aritmética \bar{x} , es decir, todas las observaciones están concentradas en un mismo punto, por lo que la dispersión es mínima (nula).

Señalar que la varianza no se suele utilizar como medida de representatividad de la media aritmética por estar expresada en unidades al cuadrado. Para ello, se utiliza la desviación típica (raíz cuadrada positiva de la varianza).

$$\text{Desviación típica: } s = +\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{N}} = \sqrt{a_2 - (a_1)^2}$$

$$\text{Desviación media: } DM_{\bar{x}} = \frac{\sum_{i=1}^8 |x_i - \bar{x}| \cdot n_i}{N} = \frac{58,3}{40} = 1,4575$$

Las medidas de dispersión utilizadas hasta ahora vienen expresadas en números concretos (unidades en las que viene medida la variable); por tanto, no son útiles para los casos en que deseamos establecer una comparación entre las dispersiones de dos muestras que vengan expresadas en distintas unidades.

En este caso hay que recurrir a medias de dispersión en números abstractos, independientes de la heterogeneidad de las unidades observadas.

- Coeficiente de Variación de Pearson: $C.V = \frac{S}{\bar{x}}$

Adviértase que este coeficiente no tiene sentido cuando $\bar{x} = 0$. A veces se multiplica por 100, para mayor comodidad en el manejo de las cifras, trabajando en porcentajes.

en nuestro caso, $C.V = \frac{s}{\bar{x}} = \frac{1,76}{4,05} = 0,4346$ (43,46%)

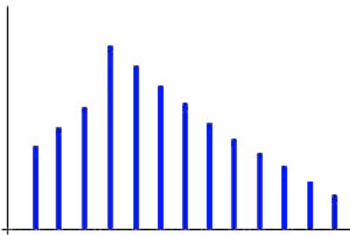
- Coeficiente de Variación Media: $CVM_{\bar{x}} = \frac{DM_{\bar{x}}}{|\bar{x}|} = \frac{1,4575}{4,05} = 0,36$

Señalar que no tiene sentido cuando $\bar{x} = 0$, o bien cuando es negativo.

MEDIDAS DE ASIMETRÍA Y APUNTAMIENTO

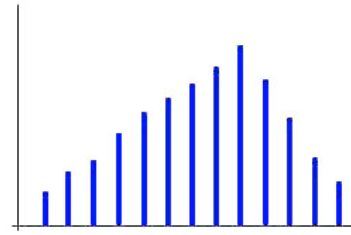
Una distribución de frecuencias es simétrica cuando los valores de la variable equidistantes de un valor central tienen las mismas frecuencias. En este caso,

$\bar{x} = M_e = M_d$. Las distribuciones que no son simétricas presentan una asimetría a la derecha o a la izquierda.



ASIMETRÍA DERECHA O POSITIVA

Desciende más lentamente por la derecha que por la izquierda: $\bar{x} \geq M_e \geq M_d$



ASIMETRÍA IZQUIERDA O NEGATIVA

Desciende más lentamente por la izquierda que por la derecha: $\bar{x} \leq M_e \leq M_d$

COEFICIENTE DE ASIMETRÍA DE PEARSON

$$A_p = \frac{\bar{x} - M_d}{s} \begin{cases} A_p > 0 & \text{Asimetría a la derecha o positiva} \\ A_p = 0 & \text{Simetría} \\ A_p < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

Este coeficiente tiene sentido cuando la moda es única

COEFICIENTE DE ASIMETRÍA DE FISHER:

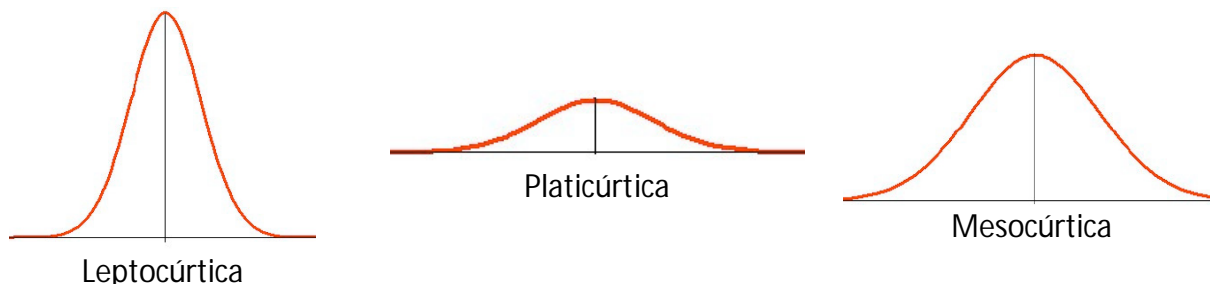
$$g_1 = \frac{m_3}{s^3} \begin{cases} g_1 > 0 & \text{Asimetría a la derecha o positiva} \\ g_1 = 0 & \text{Simetría} \\ g_1 < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

COEF. ASIMETRÍA DE BOWLEY:
(basado en la posición de los cuartiles y la mediana)

$$A_B = \frac{Q_3 + Q_1 - 2M_e}{Q_3 + Q_1} \begin{cases} A_B > 0 & \text{Asimetría a la derecha o positiva} \\ A_B = 0 & \text{Simetría} \\ A_B < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

APUNTAMIENTO O CURTOSIS

La curtosis de una distribución de frecuencias es el apuntamiento que presenta el polígono de frecuencias alrededor de la media. Si está muy apuntado diremos que la distribución es **Leptocúrtica**, si poco apuntado **Platicúrtica**, y si el apuntamiento es intermedio **Mesocúrtica** (igual apuntamiento que la normal).



COEFICIENTE DE APUNTAMIENTO O DE CURTOSIS:

$$g_2 = \frac{m_4}{s^4} - 3 \begin{cases} g_2 > 0 \rightarrow \text{LEPTOCÚRTICA (más apuntamiento que la normal)} \\ g_2 = 0 \rightarrow \text{MESOCÚRTICA (igual apuntamiento que la normal)} \\ g_2 < 0 \rightarrow \text{PLATICÚRTICA (menor apuntamiento que la normal)} \end{cases}$$

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$	$(x_i - \bar{x})^4 \cdot n_i$
1	2	2	-3,05	9,303	18,605	-56,745	173,073
2	6	12	-2,05	4,203	25,215	-51,691	105,966
3	10	30	-1,05	1,103	11,025	-11,576	12,155
4	5	20	-0,05	0,002	0,012	-0,001	0
5	10	50	0,95	0,903	9,025	8,574	8,145
6	3	18	1,95	3,803	11,408	22,245	43,377
7	2	14	2,95	8,703	17,405	51,345	151,467
8	2	16	3,95	15,603	31,205	123,260	486,876
	40	162			123,9	85,410	981,059

$$m_2 = s^2 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{123,9}{40} = 3,0975 \text{ (varianza)} \quad s = \sqrt{3,0975} = 1,76 \text{ (desviación típica)}$$

$$m_3 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^3 \cdot n_i}{N} = \frac{85,41}{40} = 2,135 \quad m_4 = \frac{\sum_{i=1}^8 (x_i - \bar{x})^4 \cdot n_i}{N} = \frac{981,059}{40} = 24,526$$

- Coeficiente de asimetría de Pearson: No tiene sentido porque no hay única moda.
- Coeficiente de asimetría de Fisher: $g_1 = \frac{m_3}{s^3} = \frac{2,135}{(1,76)^3} = 0,39 > 0 \rightarrow$ asimetría a la derecha
- Coeficiente de apuntamiento o curtosis: $g_2 = \frac{m_4}{s^4} - 3 = \frac{24,526}{(1,76)^4} - 3 = -0,44 < 0 \rightarrow$ Platicúrtica

TRANSFORMACIONES LINEALES DE LAS VARIABLES

Sea X es una variable estadística con distribución de frecuencias (x_i, n_i) , se entiende que otra variable Y es una transformación lineal de X si su distribución de frecuencias es (y_i, n_i) con $y_i = a + b \cdot x_i$ para algún par de números **a** (cambio de origen) y **b** (cambio de escala).

- Los promedios y medidas de posición son valores de la variable y , por tanto, una transformación lineal en la variable les afecta en la misma medida que a esta, ya que las frecuencias que corresponden a cada valor y y a su transformado son las mismas.

Variable	Media	Mediana	Moda	Cuantiles ($b > 0$)
$X, (x_i, n_i)$	\bar{x}	M_{ex}	M_{dx}	D_{ix}, Q_{ix}, P_{ix}
$Y, (y_i, n_i)$ $y_i = a + b \cdot x_i$	$\bar{y} = a + b \cdot \bar{x}$	$M_{ey} = a + b \cdot M_{ex}$	$M_{dy} = a + b \cdot M_{dx}$	$D_{iy} = a + b \cdot D_{ix}$ $Q_{iy} = a + b \cdot Q_{ix}$ $P_{iy} = a + b \cdot P_{ix}$

- Varianza y desviación típica se encuentran afectadas por el cambio de escala (**b**) y no por un cambio de origen (**a**)

Variable	Varianza	Desviación típica
$X, (x_i, n_i)$	s_x^2	s_x
$Y, (y_i, n_i)$ $y_i = a + b \cdot x_i$	$s_y^2 = b^2 \cdot s_x^2$	$s_y = b \cdot s_x$

- El coeficiente de variación de Pearson se encuentra afectado por un cambio de origen (**a**) pero no por un cambio de escala (**b**).

Variable	Varianza	Desviación típica	Coefficiente Variación Pearson
$X, (x_i, n_i)$	s_x^2	s_x	$CV_x = \frac{s_x}{\bar{x}}$
$Y, (y_i, n_i)$ $y_i = a + b \cdot x_i$	$s_y^2 = b^2 \cdot s_x^2$	$s_y = b \cdot s_x$	$CV_y = \frac{b \cdot s_x}{a + b \cdot \bar{x}}$
$Y, (y_i, n_i)$ $y_i = b \cdot x_i$	$s_y^2 = b^2 \cdot s_x^2$	$s_y = b \cdot s_x$	$CV_y = \frac{b \cdot s_x}{b \cdot \bar{x}} = \frac{s_x}{\bar{x}} = CV_x$

- Los coeficientes de asimetría y de curtosis permanecen invariantes ante un cambio de origen (**a**) y de escala (**b**).

Variable	Moda	Desviación típica	Asimetría Pearson	Asimetría Fisher	C. Curtosis
$X, (x_i, n_i)$	M_{dx}	s_x	A_{px}	g_{1x}	g_{2x}
$Y, (y_i, n_i)$ $y_i = a + b \cdot x_i$	$M_{dy} = a + b \cdot M_{dx}$	$s_y = b \cdot s_x$	A_{px}	g_{1x}	g_{2x}

MEDIA ARITMÉTICA Y VARIANZA DE k GRUPOS

Dados k Grupos, respectivamente, con (n_1, n_2, \dots, n_k) observaciones, medias aritméticas $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$, y varianzas $(s_1^2, s_2^2, \dots, s_k^2)$, con $N = n_1 + n_2 + \dots + n_k$. Se demuestra que:

- La media de los k Grupos:
$$\bar{X} = \frac{\sum_{i=1}^k n_i \bar{X}_i}{N}$$

Cuando en un conjunto de valores se pueden obtener dos ó más subconjuntos disjuntos, la media aritmética del conjunto se relaciona con la media aritmética de cada uno de los subconjuntos disjuntos de la siguiente forma:

$$\bar{X} = \frac{\sum_{i=1}^k \bar{X}_i n_i}{N}$$

Sea la distribución $\overbrace{x_1, x_2, \dots, x_{n_1}}^{1^\circ \text{ Grupo}}, \overbrace{x_{n_1+1}, \dots, x_{n_2}}^{2^\circ \text{ Grupo}}$, observando que habría dos subconjuntos de n_1 y $(n_2 - n_{1+1})$ elementos cada uno.

La distribución:
$$\bar{X} = \frac{\sum_{i=1}^{n_2} \bar{X}_i n_i}{N} = \frac{\sum_{i=1}^{n_1} x_i n_i + \sum_{j=n_1+1}^{n_2} x_j n_j}{N} = \frac{\sum_{i=1}^{n_1} x_i n_i}{N} + \frac{\sum_{j=n_1+1}^{n_2} x_j n_j}{N}$$

Multiplicando el numerador y denominador del primer sumando se multiplica por n_1 y el segundo por n_2

$$\bar{X} = \frac{n_1 \cdot \sum_{i=1}^{n_1} x_i n_i}{n_1 \cdot N} + \frac{n_2 \cdot \sum_{j=n_1+1}^{n_2} x_j n_j}{n_2 \cdot N} = \frac{n_1 \left(\frac{\sum_{i=1}^{n_1} x_i n_i}{n_1} \right)}{N} + \frac{n_2 \left(\frac{\sum_{j=n_1+1}^{n_2} x_j n_j}{n_2} \right)}{N} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{N}$$

- La varianza total de los k Grupos es igual a la media ponderada de las varianzas parciales más la varianza ponderada de las medias parciales:

$$\underbrace{s_x^2}_{\text{varianza total}} = \underbrace{\frac{\sum_{i=1}^k s_i^2 n_i}{N}}_{\text{media ponderada de las varianzas parciales}} + \underbrace{\frac{\sum_{i=1}^k (\bar{X}_i - \bar{X})^2 n_i}{N}}_{\text{varianza ponderada de las medias parciales}}$$

intra-grupos
entre-grupos

Mediante x_{ij} se denota en el grupo i -ésimo ($i = 1, 2, \dots, k$), la observación j -ésima ($j = 1, 2, \dots, n_i$)

$$\begin{aligned}
s^2 &= \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})]^2 = \\
&= \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 - 2 \sum_{i=1}^k (\bar{x}_i - \bar{x}) \underbrace{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)}_{=0} + \frac{1}{N} \sum_{i=1}^k (\bar{x}_i - \bar{x}) n_i = \\
&= \frac{1}{N} \sum_{i=1}^k n_i s_i^2 + \frac{1}{N} \sum_{i=1}^k (\bar{x}_i - \bar{x}) n_i
\end{aligned}$$

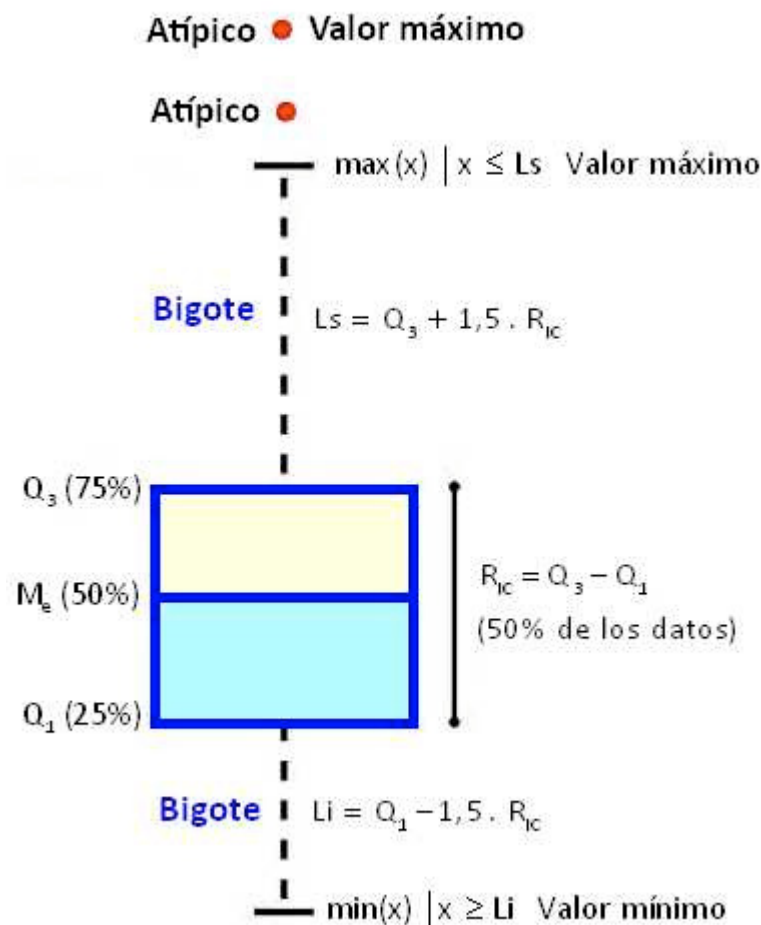
Estas propiedades adquieren un sentido importante en la técnica de Análisis de la Varianza (ANOVA)

DIAGRAMA DE CAJA Y BIGOTES

Una gráfica de este tipo consiste en una **Caja** rectangular, donde los lados más largos muestran el **recorrido intercuartílico**. Este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero (recordemos que el segundo cuartil coincide con la mediana).

Esta caja se ubica a escala sobre un segmento que tiene como extremos los valores mínimo y máximo de la variable.

Las líneas que sobresalen de la caja se llaman **Bigotes**. Estos bigotes tienen un límite de prolongación, de modo que cualquier dato o caso que no se encuentre dentro de este rango es marcado e identificado individualmente.



La Caja contiene el 50% de las observaciones centrales y su altura (base sí se coloca horizontalmente) es el recorrido intercuartílico.

El intervalo $[Li - Ls] \rightarrow \begin{matrix} Li = Q_1 - 1,5 \cdot R_{Ic} \\ Ls = Q_3 + 1,5 \cdot R_{Ic} \end{matrix}$ es el **intervalo de valores admisibles** y mide cuatro veces el recorrido intercuartílico. Los valores que quedan fuera del mismo son los que se consideran **atípicos**.

Para analizar la simetría o asimetría de un conjunto de datos a partir de este gráfico se utilizan los siguientes criterios:

- Si la línea de la Mediana está en el centro de la caja o cerca del mismo, constituye un indicio de simetría.
- Si la línea que parte de Q_3 es, aproximadamente de la misma altura que la que parte de Q_1 , también es un indicio de simetría.
- Si la línea de la Mediana se encuentra más cerca de Q_1 que del centro de la caja, es indicio de que los datos son asimétricos a la derecha.
- Si la línea que parte de Q_3 es considerablemente más larga que la que parte de Q_1 , es un indicio de simetría a la derecha o positiva.
- Son indicios de asimetría negativa que la línea de la Mediana esté más cerca de la línea de Q_3 que del centro de la caja y que la línea que parte de Q_3 sea considerablemente más corta que la que hace de Q_1

Sea la distribución X:

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

Para calcular los parámetros estadísticos, lo primero es ordenar la distribución:

x_i	20	23	24	25	29	31	33	34	36	37	39	40	41	45
n_i	1	1	3	1	1	2	1	1	2	1	2	2	1	1
N_i	1	2	5	6	7	9	10	11	13	14	16	18	19	20

5

10

15

$$Q_1 = 24$$

$$Q_2 = M_e = 33$$

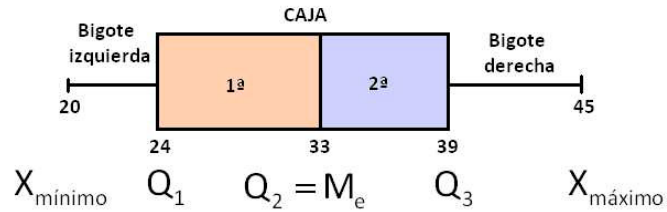
$$Q_3 = 39$$

Como $N = 20$ resulta que $20/4 = 5 \rightarrow Q_1 = 24$

$20/2 = 10 \rightarrow Q_2 = M_e = 33$

$3 \cdot 20 / 4 = 15 \rightarrow Q_3 = 39$

DIAGRAMA DE CAJA Y BIGOTES:



La mayor utilidad de los diagramas *Caja-Bigotes* es para comparar dos o más distribuciones.

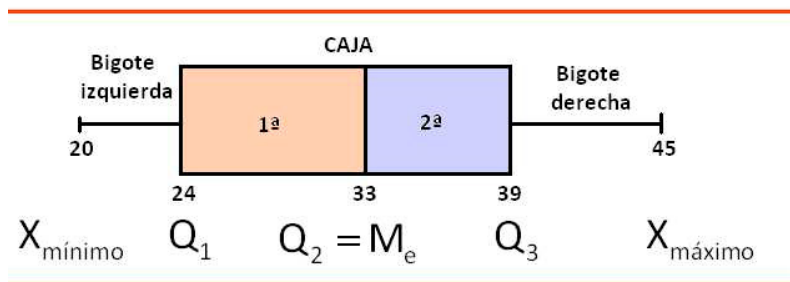
Sea la distribución Y:

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

Ordenando los datos de menor a mayor:

y_j	19	22	24	26	27	28	29	30	32	34	35	38	39	40	41	48
n_j	1	1	2	1	1	2	2	1	1	1	1	1	1	1	1	2
N_j	1	2	4	5	6	8	10	11	12	13	14	15	16	17	18	20

5
10
15
 $Q_1 = 26$
 $Q_2 = M_e = 29$
 $Q_3 = 38$



A partir de dicha comparación puede obtenerse bastante información de ambas distribuciones.

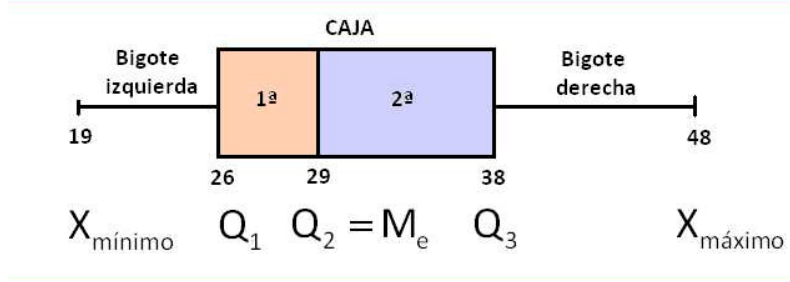




DIAGRAMA DE TALLO Y HOJA

El diagrama "tallos y hojas" (**Stem-and-Leaf Diagram**) permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica.

Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la **Hoja**) del bloque de cifras restantes (que formará el **Tallo**).

Esta representación de los datos es semejante a la de un histograma pero además de ser fáciles de elaborar, presentan más información que estos.

Ejemplo.- Horario de trenes confeccionado a partir de un díptico de la línea Castelldefels-Barcelona/Sants recogido en la estación de Renfe. Originalmente el horario ocupa una tabla de 10 filas y 9 columnas más una columna "viuda" con el tren de las 22:38. Los datos vienen dados en formato horas.minutos

5.03	7.32	9.02	11.07	13.32	15.07	16.50	18.32	20.07	22.38
6.02	7.37	9.07	11.32	13.37	15.20	17.02	18.37	20.20	
6.18	7.50	9.24	11.37	13.50	15.32	17.07	18.50	20.32	
6.37	8.02	9.32	12.02	14.02	15.37	17.20	19.02	20.37	
6.48	8.05	9.37	12.07	14.07	15.50	17.32	19.07	20.50	
6.55	8.20	10.02	12.32	14.20	16.02	17.37	19.20	21.02	
7.02	8.24	10.07	12.37	14.32	16.07	17.50	19.32	21.07	
7.07	8.32	10.32	13.02	14.37	16.20	18.02	19.37	21.20	
7.20	8.37	10.37	13.07	14.50	16.32	18.07	19.50	21.32	
7.25	8.51	11.02	13.20	15.02	16.37	18.20	20.02	21.37	

En el diagrama *Stem & Leaf* se representa la hora a la izquierda de la barra de separación | y los minutos de la salida de cada tren a la derecha.

La frecuencia de los trenes se deduce fácilmente de la longitud de las filas y es, además, muy fácil ver en que minutos de cada hora pasan típicamente los mismos.

05 03
06 02 18 37 48 55
07 02 07 20 25 32 37 50
08 02 05 20 24 32 37 51
09 02 07 24 32 37
10 02 07 32 37
11 02 07 32 37
12 02 07 32 37
13 02 07 20 32 37 50
14 02 07 20 32 37 50
15 02 07 20 32 37 50
16 02 07 20 32 37 50
17 02 07 20 32 37 50
18 02 07 20 32 37 50
19 02 07 20 32 37 50
20 02 07 20 32 37 50
21 02 07 20 32 37
22 38

Por otra parte, dado que a algunas horas se repite exactamente el horario de los trenes se puede reducir aún más el tamaño del gráfico, sin perder información y ganando en claridad.

Diagrama *Stem & Leaf* reducido:

```

          05 | 03
          06 | 02 18 37 48 55
          07 | 02 07 20 25 32 37 50
          08 | 02 05 20 24 32 37 51
          09 | 02 07 24 32 37
    10 11 12 | 02 07 32 37
13 14 15 16 17 18 19 20 | 02 07 20 32 37 50
          21 | 02 07 20 32 37
          22 | 38
  
```

Originalmente el horario ocupa una tabla de 10 filas y 9 columnas más una columna "viuda" con el tren de las 22:38. Un total de 91 campos con formato hh.mm cada uno, 455 caracteres. Al final tenemos 59 campos de 2 dígitos, 118 caracteres más los separadores, es decir 4 veces menos dígitos que con el horario original, menos espacio y más claridad.

Esto da idea de que una disposición apropiada de los datos puede ser doblemente informativa y que la representación gráfica puede contribuir enormemente a la percepción de patrones y a la comprensión de la naturaleza de los fenómenos.

Ejemplo.- Sea la distribución de frecuencias:

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

Ordenando los datos de menor a mayor:

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

- Se comienza seleccionando el **tallo**: cifras de decenas (3, 2, 4), que reordenadas son 2, 3 y 4.
- Se «añade» cada **Hoja** a su **Tallo** (unidades)

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

Tallo decenas	Hoja unidades							
2	0	3	4	4	4	5	9	
3	1	1	3	4	6	6	7	9 9
4	0	0	1	5				

 Para comparar dos distribuciones, sea otra distribución de datos:

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

Ordenando los datos de menor a mayor:

Hoja (N=20) unidades	Tallo decenas	Hoja (N=20) unidades
	1	
9 9 8 8 7 6 4 4 2	2	0 3 4 4 4 5 9
	3	1 1 3 4 6 6 7 9 9
	4	0 0 1 5

Ejemplo.- El tratamiento de los niños con desórdenes de la conducta puede ser complejo. Además del reto que ofrece el tratamiento, se encuentra la falta de cooperación del niño/niña y el miedo y la falta de confianza de los adultos. Para diseñar el tratamiento un psiquiatra considero una muestra aleatoria de 20 niños, anotando el tiempo necesario que requiere en cada niño para lograr un plan integral, los resultados obtenidos en horas son:

6 7 7 8 8 8 8 9 9 9 9 9 9 9 10 10 10 10 10 11

CONSTRUCCIÓN DIAGRAMA DE CAJA:

x_i	6	7	8	9	10	11
n_i	1	2	4	7	5	1
N_i	1	3	7	14	19	20

$5 \quad 10 \quad 15$

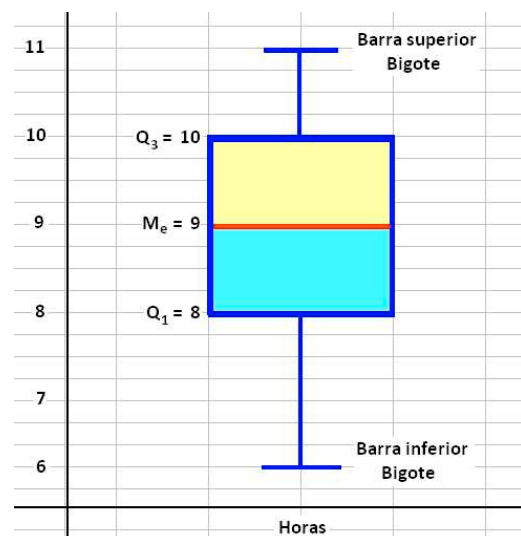
$$20/4 = 5 \rightarrow Q_1 = 8 \quad 20/2 = 10 \rightarrow Q_2 = M_e = 9 \quad 3 \cdot 20 / 4 = 15 \rightarrow Q_3 = 10$$

$$\text{Rango} = 11 - 6 = 5 \text{ horas} \quad R_{1c} = Q_3 - Q_1 = 10 - 8 = 2 \text{ horas}$$

$$\text{Intervalo de valores admisibles: } [Li - Ls] \rightarrow \begin{cases} Li = Q_1 - 1,5 \cdot R_{1c} \rightarrow Li = 8 - 1,5 \cdot 2 = 5 \\ Ls = Q_3 + 1,5 \cdot R_{1c} \rightarrow Ls = 10 + 1,5 \cdot 2 = 13 \end{cases}$$

Valores extremos $\begin{cases} \exists x_i / x_i < 5 = Li \rightarrow 6 > 5 \text{ Por tanto 6 no es valor extremo inferior} \\ \exists x_i / x_i > 13 = Ls \rightarrow 11 < 13 \text{ Por tanto 11 no es valor extremo superior} \end{cases}$

La caja muestra cierta simetría, aunque los bigotes dicen lo contrario, mostrando un sesgo a la izquierda.



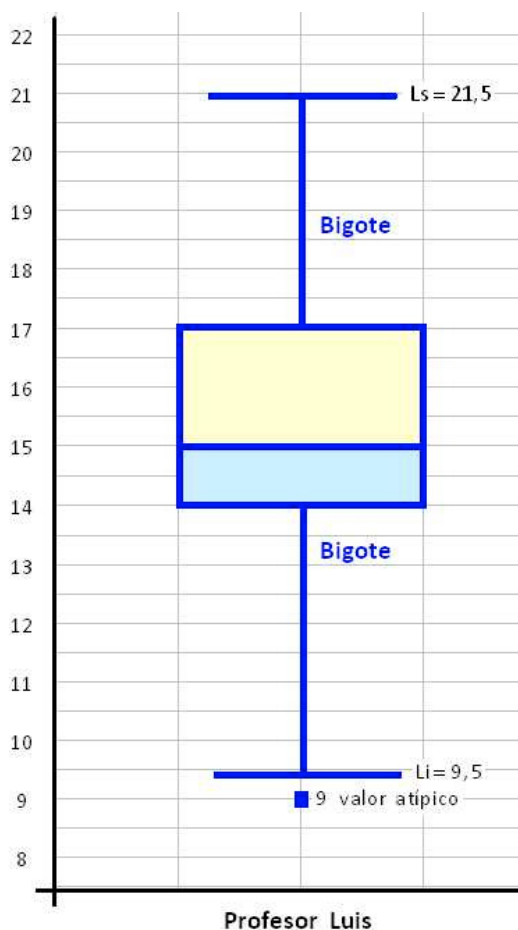
CONSTRUCCIÓN DIAGRAMA TALLO Y HOJA (Stem & Leaf):

6 7 7 8 8 8 8 9 9 9 9 9 9 9 10 10 10 10 10 11

Tallo Stem	Hoja Leaf
6	0
7	0 0
8	0 0 0 0
9	0 0 0 0 0 0 0
10	0 0 0 0 0
11	0

La distribución no es simétrica con un leve sesgo a la izquierda.

Ejemplo.- Dos profesores (Luis y Miguel) están interesados en estudiar los hábitos de sueño de los estudiantes. Para ello, registran el tiempo (en minutos) que demoran en quedarse dormidos sus alumnos desde que comienza la clase. El gráfico muestra los tiempos que tardan en quedarse dormidos los alumnos del profesor Luis.



$$\text{Rango} = 21 - 9 = 12 \text{ minutos}$$

$$Q_1 = 14 \quad M_e = 15 \quad Q_3 = 17$$

$$R_{Ic} = Q_3 - Q_1 = 17 - 14 = 3 \text{ minutos}$$

$$[Li - Ls] \rightarrow \begin{aligned} Li &= Q_1 - 1,5 \cdot R_{Ic} \rightarrow Li = 14 - 1,5 \cdot 3 = 9,5 \\ Ls &= Q_3 + 1,5 \cdot R_{Ic} \rightarrow Ls = 17 + 1,5 \cdot 3 = 21,5 \end{aligned}$$

Intervalo de valores admisibles: $[9,5 - 21,5]$

Los bigotes son los segmentos verticales que parten de la caja y llegan hasta el menor y mayor valor observado que sea admisible.

Por debajo de la caja, se encuentra el valor atípico 9 (fuera del menor valor admisible).

La caja presenta una asimetría a la derecha.

Los datos del profesor Miguel son los siguientes:

10,5 11,3 11,9 12 12,3 12,3 12,5 12,7 13,4 13,7
13,8 14,2 14,8 15,1 15,3 16,7 16,8 18,8 20,8

Para construir un diagrama de caja:

Posición de la Mediana: $\frac{N+1}{2} = \frac{20}{2} = 10 \rightarrow M_e = 13,7$

Posición Q_1 : $\frac{N+1}{4} = \frac{20}{4} = 5 \rightarrow Q_1 = 12,3$

Posición Q_3 : $\frac{3N+1}{4} = \frac{58}{4} = 14,5 \rightarrow Q_3 = 15,3$

$R_{IC} = Q_3 - Q_1 = 15,3 - 12,3 = 3$ minutos

Intervalo de valores admisibles: $[Li - Ls] \rightarrow$
 $Li = Q_1 - 1,5 \cdot R_{IC} \rightarrow Li = 12,3 - 1,5 \cdot 3 = 7,8$
 $Ls = Q_3 + 1,5 \cdot R_{IC} \rightarrow Ls = 15,3 + 1,5 \cdot 3 = 19,8$

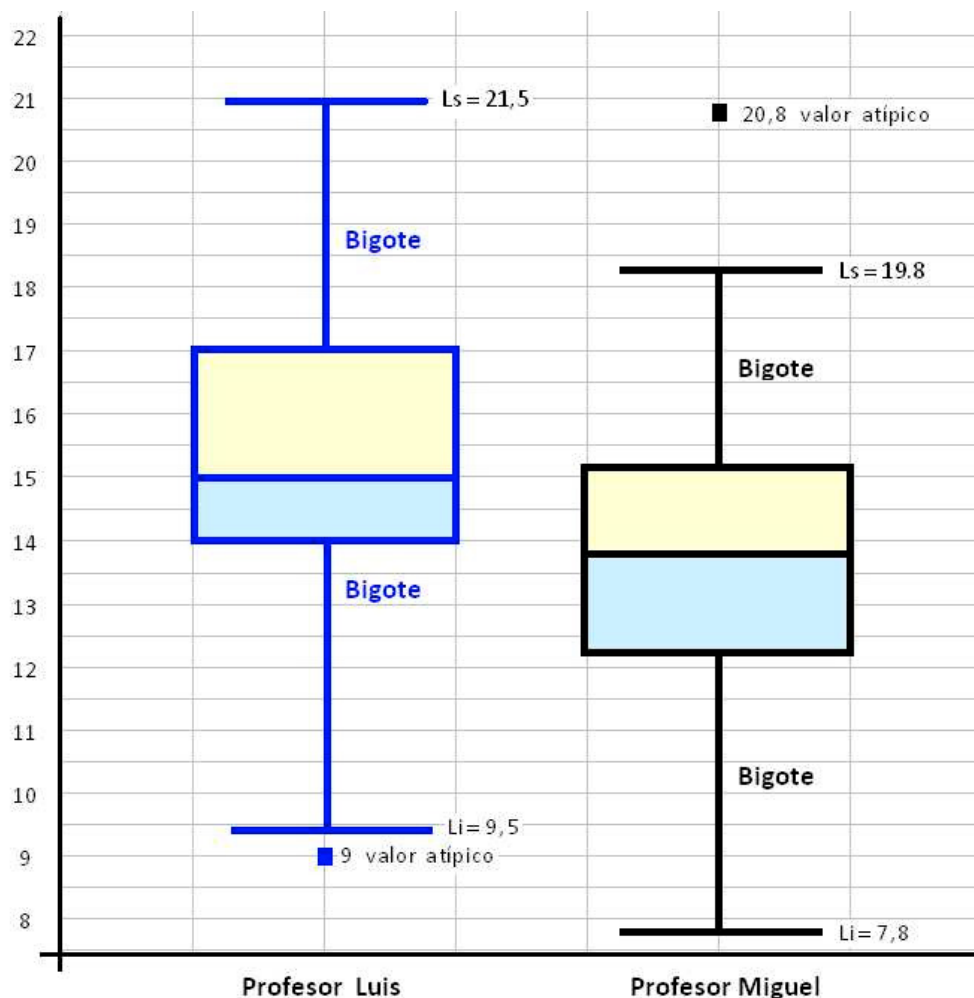
Cálculo de valores extremos:

- $\exists x_i / x_i < 7,8 = Li \rightarrow 10,5 > 7,8 \Rightarrow 10,5$ no es valor extremo inferior
- $\exists x_i / x_i > 19,8 = Ls \rightarrow 20,8 > 19,8 \Rightarrow 20,8$ es valor extremo superior

Verificando si el número anterior es valor extremo superior:

$\exists x_i / x_i > 19,8 = Ls \rightarrow 18,8 < 19,8 \Rightarrow 18,8$ no es valor extremo superior

$\exists x_i / x_i > 19,8 = Ls \rightarrow 20,8 > 19,8 \Rightarrow 20,8$ es valor extremo superior



MEDIDAS DE CONCENTRACIÓN: ÍNDICE DE GINI Y CURVA DE LORENZ

Cuando se realizaba un estudio descriptivo de los valores observados en una variable, la palabra concentración era la opuesta a dispersión.

A partir de ahora el objeto del estudio será el total de los recursos repartidos entre todos los individuos que intervienen en la distribución.

Si a cada individuo n_i se le atribuye una cantidad x_i de recursos (euros si se analizan los salarios; toneladas de carbón si se estudia la producción de carbón de una zona...), el total de recursos que se reparten N individuos o entes que forman la distribución será:
$$\sum_{i=1}^k x_i \cdot n_i = N \cdot \bar{x}$$

La cantidad total de recursos no suele siempre repartirse de forma equitativa, sino que al haber distintos valores posibles de la variable x_i habrá individuos que se repartan una mayor cantidad de recursos que otros. Es este aspecto el que se desea estudiar con las medidas de concentración.

En esta línea, se dice que una distribución está muy concentrada si la suma total de sus valores se encuentra muy concentrada en pocos individuos, mientras que se dice que está poco concentrada si sus recursos se encuentran repartidos entre sus individuos. Cuando los recursos están perfectamente distribuidos se dice que la variable está equilibrada.

Para analizar la concentración de la distribución se realiza el seguimiento de los recursos repartidos en la distribución a medida que se van asignando a los individuos. Para ello, se ordena a los individuos en orden creciente, observando como al avanzar los valores ocurridos de la variable van evolucionando el número de individuos que tienen asignada una cantidad de recursos (por un lado) y el total de recursos que esos individuos se han repartido (por otra parte).

Las mediciones se realizan asociando a cada posible valor observado de la variable x_i dos valores:

- La frecuencia relativa acumulada que le corresponde, $p_i = \frac{N_i}{N}$, es decir, el número de individuos que perciben recursos inferiores o iguales a esa cantidad.

- La proporción de recursos que llevan repartida entre ellos $q_i = \frac{\sum_{i=1}^m x_i \cdot n_i}{\sum_{i=1}^k x_i \cdot n_i} = \frac{\sum_{i=1}^m x_i \cdot n_i}{N \cdot \bar{x}} = \frac{u_i}{u_k}$

Al ir aumentando el valor de la variable sobre la distribución, si la proporción de recursos que se van repartiendo es aproximadamente igual a la proporción de individuos que se lo van repartiendo, la variable estará bien repartida y, en consecuencia, la distribución estará poco concentrada.

Si por el contrario, los individuos que menos perciben se reparten una proporción de recursos muy inferiores a la que representan ellos con respecto al total de los individuos, la distribución será muy concentrada y los recursos mal repartidos.

El análisis descrito se visualiza de forma sencilla sobre un cuadrado de lado unidad (100%) y recibe el nombre de curva de concentración de Lorenz.

Para el desarrollo del estudio se parte de una tabla como la que se describe:

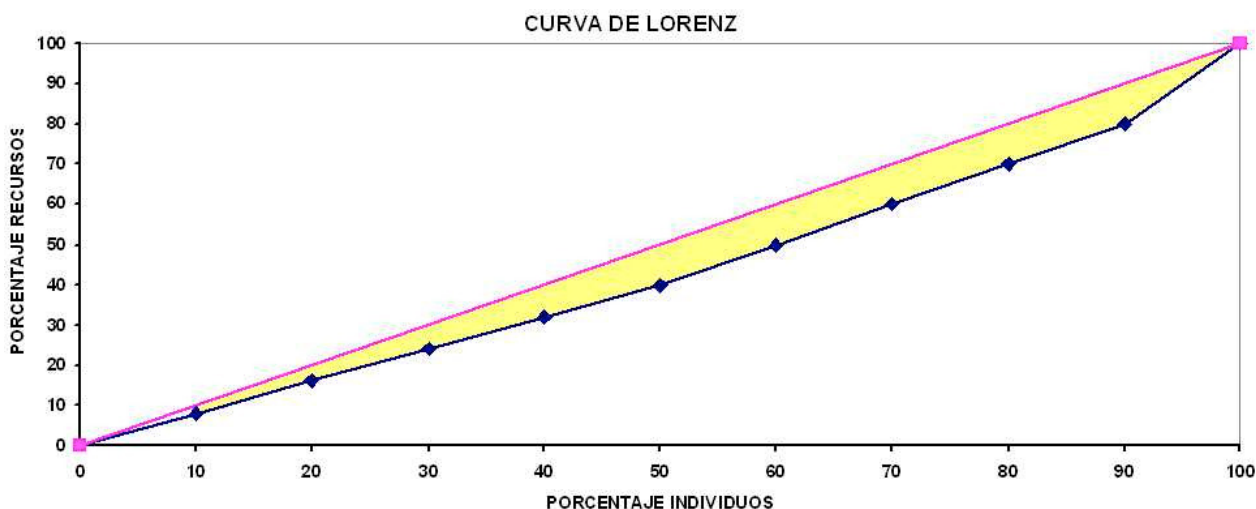
x_i	n_i	$x_i \cdot n_i$	N_i	$u_i = x_i \cdot n_i$ acumulada	$\% p_i = \frac{N_i}{N} \cdot 100$	$\% q_i = \frac{u_i}{u_k} \cdot 100$
x_1	n_1	$x_1 \cdot n_1$	N_1	$u_1 = x_1 \cdot n_1$	p_1	q_1
x_2	n_2	$x_2 \cdot n_2$	N_2	$u_2 = x_1 \cdot n_1 + x_2 \cdot n_2$	p_2	q_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_i	$x_i \cdot n_i$	N_i	$u_i = \sum_{i=1}^i x_i \cdot n_i$	p_i	q_i
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_k	n_k	$x_k \cdot n_k$	N_k	$u_k = \sum_{i=1}^k x_i \cdot n_i$	p_k	q_k
	N	$\sum_{i=1}^k x_i \cdot n_i$				

En la columna $u_i = x_i \cdot n_i$ acumulada se expresan las sumas acumuladas parciales del recurso hasta el lugar correspondiente, y donde p_i y q_i son las proporciones de individuos y recursos acumulados, respectivamente, que se pueden calcular:

$$p_i = \frac{N_i}{N} \quad q_i = \frac{u_i}{u_k} \quad (.100 \text{ expresados en porcentajes})$$

Al estar ordenados los x_i en orden creciente, al principio aparecen los que menos perciben y, por tanto, la proporción de individuos siempre tiene que avanzar más rápidamente que la proporción de recursos repartidos. Así, pues, siempre $p_i \geq q_i$

Para la curva de Lorenz construimos un cuadrado de lado unidad (100%), representando en los ejes los valores p_i (individuos) y q_i (recursos). Si sobre el eje de abscisas se representan los valores p_i (individuos) y sobre el eje de ordenadas los valores q_i (recursos), la curva formada siempre estará por debajo de la diagonal principal del cuadrado.



Es evidente que si $p_i = q_i$ la curva coincidiría con la diagonal, la proporción de individuos y de recursos irían evolucionando conjuntamente, y sería un caso de variable equidistribuida.

Si por el contrario, la curva se va alejando hacia los lados del cuadrado, dejando entre ella y la diagonal un área considerable, cuanto mayor sea esta separación, mayor será la concentración y peor el reparto de recursos.

El caso de máxima concentración se alcanzaría cuando la curva de Lorenz coincidiese con los lados del cuadrado, dejando un área entre ella y la diagonal de 0,5 (mitad del área del cuadrado de lado unidad).

La idea de medir el área entre la diagonal y la curva de Lorenz da como resultado el llamado Índice de Concentración de Gini, que viene expresado:

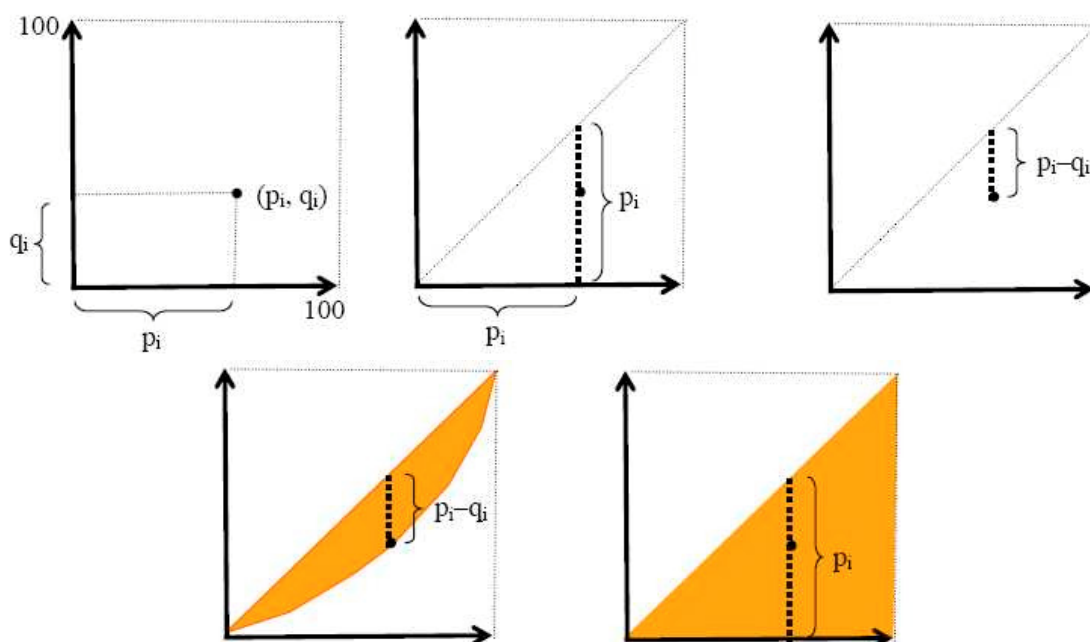
$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} \quad 0 \leq I_G \leq 1$$

$I_G = 0$	→	equidistribución
I_G	+ próximo	→ 0 (mejor reparto equitativo)
I_G	+ próximo	→ 1 (peor reparto ≡ mayor concentración)
$I_G = 1$	→	un individuo se lleva el total de recursos

Adviértase que: $\frac{\sum (p_i - q_i)}{\sum p_i} = \frac{\sum p_i - \sum q_i}{\sum p_i} = \frac{\sum p_i}{\sum p_i} - \frac{\sum q_i}{\sum p_i} = 1 - \frac{\sum q_i}{\sum p_i}$ (propiedad sumatorio)

RELACIÓN ENTRE ÍNDICE CONCENTRACIÓN DE GINI Y CURVA DE LORENZ

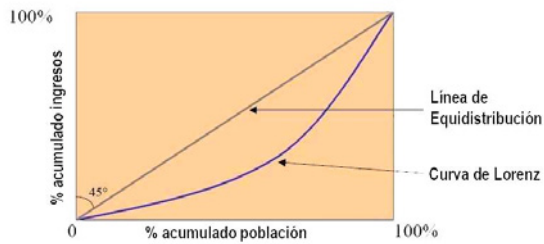
Dado un punto cualquiera (p_i, q_i) de la curva de Lorenz



- $p_i = q_i \quad \forall i$ $\xrightarrow{\text{curva de Lorenz sobre la diagonal}}$ $I_G = 0 \rightarrow$ Equidistribución
- $q_1 = q_2 = q_3 = \dots = q_{k-1} = 0$ $\xrightarrow{\text{curva de Lorenz sobre lados del cuadrado}}$ $I_G = 1 \rightarrow$ Máxima Concentración

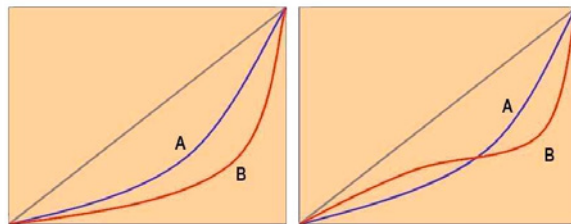
Curva de Lorenz

Muestra el porcentaje acumulado de ingreso que poseen los individuos u hogares, ordenados en forma ascendente de acuerdo con su nivel de ingreso.



Curva de Lorenz

Para determinar el grado de desigualdad, se compara las Curvas de Lorenz.

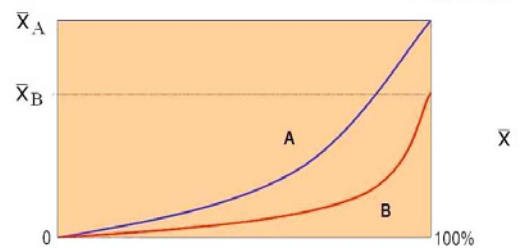


A "domina" a B \Rightarrow
 \Rightarrow Desigualdad es menor en A

A v B se cruzan \Rightarrow
 \Rightarrow No es posible establecer comparaciones

Curva de Lorenz Generalizada

En caso de que las Curvas de Lorenz se crucen, es posible utilizar la CL Generalizada, multiplicando los valores por la media de cada distribución (\bar{x}_A y \bar{x}_B).



El coeficiente de Corrado Gini (1884-1965) satisface **cuatro principios para medir la desigualdad**:

- ◆ **Principio de anonimato.**- Si se produce una modificación en una distribución de renta consistente en que dos individuos intercambien sus rentas, el valor del índice no debe variar.
- ◆ **Principio de la población.**- Si se multiplica por un mismo escalar el tamaño de todos los conjuntos de individuos con la misma renta, el valor del índice no debe variar. Es decir, el tamaño de la población no importa, lo que interesa son las proporciones de individuos de la población que perciben diferentes niveles de renta.
- ◆ **Principio de la renta relativa.**- El índice debe mantenerse invariante frente a las variaciones proporcionales en todas las rentas.
- ◆ **Principio de Dalton.**- Toda transferencia de renta de un individuo a otro más rico ha de *aumentar* el valor de la desigualdad, y recíprocamente toda transferencia de renta de un individuo a otro más pobre ha de *reducir* el índice, siempre que la ordenación relativa de los individuos se mantenga.

MEDIDAS DE CONCENTRACIÓN INDUSTRIAL

ÍNDICE DE HERFINDAHL

Sea un conjunto de k empresas con una cifra de ventas x_i , se denomina Cuota de Mercado a la

cantidad:
$$s_i = \frac{x_i}{\sum_{i=1}^k x_i}$$

El Índice de Concentración de Herfindahl es la cantidad:
$$H = \sum_{i=1}^k s_i^2$$

PROPIEDADES DEL ÍNDICE DE HERFINDAHL:

- Es un índice acotado: $\frac{1}{k} \leq H \leq 1$
- En un mercado de competencia perfecta: $H = \frac{1}{k}$, puesto que $H = \sum_{i=1}^k \left[\frac{1}{k} \right]^2 = \frac{1}{k}$
- En un mercado monopolístico: $H = 1$

Ejemplo: Importaciones de vino (millones de litros) según el Banco Mundial en 2004:

Francia	Italia	Australia	España	Chile	Alemania	Portugal
1049	952	731	183	152	80,6	53

Cuota de mercado e Índice de Herfindahl:

	Francia	Italia	Australia	España	Chile	Alemania	Portugal	
	1049	952	731	183	152	80,6	53	3204,6
s_i	0,3273	0,2971	0,2281	0,0571	0,0474	0,0252	0,0178	1
s_i^2	0,1072	0,0883	0,0520	0,0033	0,0022	0,0006	0,0003	0,2539

La cuota de mercado de cada país se calcula mediante la fórmula: $s_i = x_i / \sum_{i=1}^7 x_i$.

En el caso de competencia perfecta la cuota de mercado sería $s_i = 1/7 = 0,1428$, por lo que se puede afirmar que en este mercado no existe competencia perfecta.

El índice de concentración de Herfindahl: $H = \sum_{i=1}^7 s_i^2 = 0,2539$

Índice de Theil

Indicador de desigualdad en el reparto de una magnitud entre distintas unidades preceptoras o de asignación (el reparto puede tener lugar entre personas (rentas), empresas (cuotas de mercado), unidades espaciales (provincias o regiones), etc., fue introducido inicialmente como una medida de entropía dentro del contexto de la teoría de la información.

La entropía sirve para medir el grado de desorden en un sistema (un sistema desordenado sería equivalente a otro en el que cada uno de los componentes del mismo no están 'equilibrados') y también para comparar situaciones distintas.

El índice de Theil se define, inicialmente, en términos de las probabilidades de los distintos valores de una distribución. Sin embargo, esas probabilidades pueden aproximarse por las frecuencias relativas observadas para esos valores o simplemente por un conjunto de proporciones, con las únicas condiciones de que sean no negativas y que su suma sea igual a la unidad.

En el marco de este contexto, sean N individuos con rentas (x_1, x_2, \dots, x_k) , la proporción de la masa

total de las rentas que corresponde al individuo i -ésimo será:
$$p_i = \frac{x_i}{\sum_{i=1}^k x_i \cdot n_i}$$

El índice de Theil, se define: $T = \ln N - H_N(p_i)$

se basa en la entropía o medida del desorden:
$$H_N(p_i) = \sum_{i=1}^k p_i \cdot n_i \cdot \ln \left[\frac{1}{p_i} \right] = - \sum_{i=1}^k p_i \cdot n_i \cdot \ln(p_i)$$

con lo que, $T = \ln N + \sum_{i=1}^k p_i \cdot n_i \cdot \ln(p_i)$.

El índice de Theil relativo: $\bar{T} = \frac{T}{\ln N} \quad 0 \leq \bar{T} \leq 1$

El índice de Theil presenta el inconveniente de que depende del número máximo de observaciones. Los valores extremos del índice de Theil son $[0, \ln N]$

- $T = 0$ cuando $[p_1 = p_2 = \dots = p_{k-1} = 0]$ (toda la cuota de mercado de todas las empresas p_i vale cero, salvo la de una que es la unidad. Situación de monopolio). La concentración es máxima.
- $T = \ln N$ cuando $[p_1 = p_2 = \dots = p_k]$ (la cuota de mercado de todas las empresas p_i son iguales. Situación de reparto igualitario). La concentración es mínima.
- $\begin{cases} T \longrightarrow \ln N \mapsto \text{Mayor equidad (mínima concentración)} \\ T \longrightarrow 0 \mapsto \text{Menor equidad (máxima concentración)} \end{cases}$
- El índice de Theil no requiere la ordenación de los valores
- Si algún $p_i = 0 \mapsto p_i \cdot n_i \cdot \ln(p_i) = 0$
- El índice de Theil permite descomponer la desigualdad en subgrupos, en este sentido se puede estudiar la desigualdad debida a cada uno de los subgrupos.

Sea X la variable observada, suponiendo que X se agrupa en k grupos (G_1, G_2, \dots, G_k) , respectivamente, de tamaños (N_1, N_2, \dots, N_k) .

donde,
$$p_g = \frac{\sum_{i \in G} x_i}{\sum_{i=1}^k x_i \cdot n_i} \quad g = 1, \dots, k$$

En cada uno de los grupos: $T_g = \ln N_g + \sum_{i=1}^{N_g} p_i \cdot \ln \left[\frac{1}{p_i} \right]$

$$\text{Entonces, } T = \overbrace{\ln N + \sum_{g=1}^k p_g \cdot \ln \left[\frac{p_g}{N_g} \right]}^{\text{Desigualdad entre Grupos INTERGRUPOS}} + \overbrace{\sum_{g=1}^k p_g \cdot T_g}_{\text{Desigualdad dentro grupos INTRAGRUPOS}}$$

- **DESIGUALDAD INTERGRUPOS:** $\ln N + \sum_{g=1}^k p_g \cdot \ln \left[\frac{p_g}{N_g} \right]$, mide la disparidad entre grupos teniendo en cuenta el tamaño de cada grupo N_g en relación al peso del grupo p_g en la variable económica observada.
- **DESIGUALDAD INTRAGRUPOS:** $\sum_{g=1}^k p_g \cdot T_g$ mide la disparidad dentro de los grupos, es la media de los coeficientes de Theil de cada grupo ponderados por los pesos de cada grupo.

Ejemplo 1: Dada la distribución de salarios semanales (euros), determinar el coeficiente de Theil

x_i	n_i	$x_i \cdot n_i$	$p_i = \frac{x_i}{\sum_{i=1}^k x_i \cdot n_i}$	$\ln p_i$	$p_i \cdot n_i \cdot \ln p_i$
80	10	800	0,01	-4,605	-0,461
150	20	3000	0,01875	-3,977	-1,491
200	15	3000	0,025	-3,689	-1,383
240	5	1200	0,03	-3,507	-0,526
	50	8000			-3,861

$$H_{50}(p_i) = -\sum_{i=1}^4 p_i \cdot n_i \cdot \ln(p_i) = 3,861 \quad \ln(50) = 3,912$$

$$\text{Índice de Theil : } T = \ln 50 - H_{50}(p_i) = 3,912 - 3,861 = 0,051$$

$$\text{Índice de Theil relativo: } \bar{T} = \frac{T}{\ln 50} = \frac{0,051}{3,912} = 0,013$$

Ejemplo 2: La tabla adjunta recoge datos sobre el valor añadido bruto (VAB) en u.m. de siete regiones vinícolas españolas.

Regiones vinícolas	R1	R2	R3	R4	R5	R6	R7
VAB (x_i)	2460,5	619	613,2	1150	1865	437,1	661,9

- Hallar el coeficiente de Theil
- Analizar el coeficiente de Theil mediante un análisis desagregado dividiendo las regiones vinícolas en dos grupos (R1-R4-R5 y R2-R3-R6-R7)

Solución:

a)

Regiones vinícolas	VAB (x_i)	$p_i = \frac{x_i}{\sum_{i=1}^k x_i \cdot n_i}$	$\ln p_i$	$p_i \cdot n_i \cdot \ln p_i$
R1	2460,5	0,315	-1,155	-0,364
R2	619	0,079	-2,535	-0,201
R3	613,2	0,079	-2,544	-0,200
R4	1150	0,147	-1,915	-0,282
R5	1865	0,239	-1,432	-0,342
R6	437,1	0,056	-2,883	-0,161
R7	661,9	0,085	-2,468	-0,209
	7806,7	1		-1,759

$$H_7(p_i) = -\sum_{i=1}^7 p_i \cdot n_i \cdot \ln(p_i) = 1,759 \quad \ln(7) = 1,946$$

$$\text{Índice de Theil : } T = \ln 7 - H_7(p_i) = 1,946 - 1,759 = 0,187$$

b) Dividiendo las regiones vinícolas en los grupos indicados:

Regiones vinícolas	VAB (x_i)	$p_g = \frac{\sum_{i \in G} x_i}{\sum_{i=1}^7 x_i \cdot n_i}$	$p_i = \frac{x_i}{\sum_{i=1}^4 x_i \cdot n_i}$	$\ln p_i$	$p_i \cdot n_i \cdot \ln p_i$
R2	619	0,079	0,266	-1,326	-0,352
R3	613,2	0,079	0,263	-1,335	-0,351
R6	437,1	0,056	0,188	-1,674	-0,314
R7	661,9	0,085	0,284	-1,259	-0,357
	2331,2	0,299	1		-1,375

$$T_g = \ln N_g + \sum_{i=1}^{N_g} p_i \cdot \ln \left[\frac{1}{p_i} \right] \xrightarrow{g=1 \quad N_1=4} T_1 = \ln N_1 + \sum_{i=1}^4 p_i \cdot \ln p_i = \ln 4 - 1,357 = 0,0113$$

Regiones vinícolas	VAB (x_i)	$p_g = \frac{\sum_{i \in G} x_i}{\sum_{i=1}^7 x_i \cdot n_i}$	$p_i = \frac{x_i}{\sum_{i=1}^3 x_i \cdot n_i}$	$\ln p_i$	$p_i \cdot n_i \cdot \ln p_i$
R1	2460,5	0,315	0,449	-0,800	-0,359
R4	1150	0,147	0,210	-1,561	-0,328
R5	1865	0,239	0,341	-1,077	-0,367
	5475,5	0,701	1	-3,437	-1,054

$$T_g = \ln N_g + \sum_{i=1}^{N_g} p_i \cdot \ln \left[\frac{1}{p_i} \right] \xrightarrow{g=2 \quad N_2=3} T_2 = \ln N_2 + \sum_{i=1}^3 p_i \cdot \ln p_i = \ln 3 - 1,054 = 0,0446$$

- Desigualdad **INTERGRUPOS** (entre grupos):

$$\begin{aligned} \ln N + \sum_{g=1}^k p_g \cdot \ln \left[\frac{p_g}{N_g} \right] &\equiv \ln N + \sum_{g=1}^2 p_g \cdot \ln \left[\frac{p_g}{N_g} \right] = \ln 7 + (0,299) \cdot \ln \left[\frac{0,299}{4} \right] + (0,701) \cdot \ln \left[\frac{0,701}{3} \right] = \\ &= 1,9459 - 0,7755 - 1,0192 = 0,1512 \end{aligned}$$

- Desigualdad **INTRAGRUPOS** (dentro de los grupos):

$$\sum_{g=1}^k p_g \cdot T_g \equiv \sum_{g=1}^2 p_g \cdot T_g = (0,299) \cdot (0,0113) + (0,701) \cdot (0,446) = 0,0346$$

De este modo, el coeficiente de Theil:

$$T = \ln N + \sum_{g=1}^k p_g \cdot \ln \left[\frac{p_g}{N_g} \right] + \sum_{g=1}^k p_g \cdot T_g = 0,1512 + 0,0346 = 0,1858$$

En términos relativos: $\frac{T}{0,1858} = \frac{0,1512}{0,1858} + \frac{0,0346}{0,1858} = 0,8138 + 0,1862 = 1$

De la desigualdad existente en las siete regiones vinícolas, el 81,38% es consecuencia a la desigualdad entre los grupos. A la hora de tomar medidas económicas para disminuir, aún más, la desigualdad, se actuaría en esta dirección, tratando de limar las diferencias entre los dos grupos (esta es una de las ventajas del análisis desagregado, permite determinar el origen de las diferencias existentes entre regiones, comunidades, etc.)

COEFICIENTE DE CORRELACIÓN DE SPEARMAN

A veces interesa hallar la correlación existente entre dos series de datos, en donde los valores de sus variables no vienen señalados por sus frecuencias absolutas, sino en el orden que ocupan en la observación. Para estudiar la situación descrita se utiliza el coeficiente de Spearman.

El coeficiente de correlación de Spearman r_s estudia la correlación (asociación) entre dos variables (cuando ambas son de tipo discreto, o cuando no presentan una distribución parecida a la normal), se define:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n-1) \cdot (n+1)} \quad -1 \leq r_s \leq 1$$

Para calcular el coeficiente r_s hay que ordenar los datos en función de cada valor x_i asignando rango a cada valor. Se repite la operación en función de cada valor y_i asignando rango a cada valor. En esta línea, d_i es la diferencia entre el orden obtenido por el individuo i -ésimo en ambas series de datos.

Cuando el coeficiente de correlación r_s de Spearman presenta un valor cercano a cero se dice que ambas variables no presentan correlación (asociación).

Se plantean las hipótesis $\left\{ \begin{array}{l} \text{Hipótesis nula: } H_0 : r_s = 0 \\ \text{Hipótesis alternativa: } H_a : r_s \neq 0 \end{array} \right.$ Se rechaza H_0 si $r_s \geq r_{\text{crítico}}$

La hipótesis nula H_0 (no existe correlación) se rechaza cuando el valor del coeficiente de correlación de Spearman r_s calculado supera a un valor crítico $r_{\text{crítico}}$ del coeficiente de correlación de Spearman con determinado nivel de fiabilidad (tablas).

Es decir, se rechaza la hipótesis nula H_0 (no existe correlación) cuando $r_s \geq r_{\text{crítico}}$

En caso contrario, se acepta la hipótesis alternativa, concluyendo que existe correlación entre las variables con determinado grado de fiabilidad.

Ejemplo: Con la pretensión de averiguar si existe correlación en las asignaturas de Estadística y Macroeconomía se recogen las puntuaciones (números enteros) obtenidas por diez alumnos:

		Rangos puntuaciones			
Estadística	Macroeconomía	Estadística	Macroeconomía	d_i	d_i^2
65	74	4	6	-2	4
72	61	5	3	2	4
75	69	6	5	1	1
82	90	7	9	-2	4
50	51	1	1	0	0
95	79	10	8	2	4
87	95	9	10	-1	1
53	52	2	2	0	0
83	77	8	7	1	1
64	63	3	4	-1	1
EXCEL:		=JERARQUIA(A1:A\$10;1)	=JERARQUIA(B1:B\$10;1)		20

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{10} d_i^2}{n \cdot (n-1) \cdot (n+1)} = 1 - \frac{6 \cdot 20}{10 \cdot 9 \cdot 11} = 0,88$$

El coeficiente de correlación de Spearman es alto (próximo a 1), indicando una buena asociación de tipo discreto entre ambas variables (Estadística, Macroeconomía), es decir, las más altas puntuaciones en una de las variables correspondieron a las más altas puntuaciones en la otra y, complementariamente, las más bajas puntuaciones en una variable correspondieron a las más bajas puntuaciones de la otra.

Se plantean las hipótesis $\left\{ \begin{array}{l} \text{Hipótesis nula: } H_0 : r_s = 0 \\ \text{Hipótesis alternativa: } H_a : r_s \neq 0 \end{array} \right.$ Se rechaza H_0 si $r_s \geq r_{\text{crítico}}$

Para un grupo de 10 estudiantes ($n = 10$), el valor calculado de $r_s = 0,88$, con un nivel de confianza del 95% ($p\text{-valor} = 0,05$), es superior al valor crítico de $r_{\text{crítico}} = 0,564$ ($r_s = 0,88 > r_{\text{crítico}} = 0,564$), rechazando la hipótesis nula y concluyendo que existe asociación directa entre los aciertos que obtuvieron los alumnos en las pruebas de Estadística y Macroeconomía.

Valores críticos del coeficiente r_s

	Nivel de significación	
	0,05	0,01
4	1,000	-----
5	0,900	1,000
6	0,829	0,943
7	0,714	0,893
8	0,643	0,833
9	0,600	0,783
10	0,564	0,746
12	0,506	0,712
14	0,456	0,645
16	0,425	0,601
18	0,399	0,564
20	0,377	0,534
22	0,359	0,508
24	0,343	0,485
26	0,329	0,465
28	0,317	0,448
30	0,306	0,432

