

1. Variables estadísticas unidimensionales

1.1. Conceptos básicos: población, muestra y variable estadística

Definiciones

- **Población:** Conjunto de elementos que poseen una característica sobre el que se desea realizar un estudio estadístico. El número de elementos de la población se llama tamaño poblacional. 
- **Muestra:** Subconjunto finito de una población que conserva las características de la población y la representa. Para ello la muestra debería ser aleatoria. El número de elementos de la muestra se llama tamaño muestral.
- **Variable estadística:** Característica de una población que se quiere estudiar (se denotan con letras mayúsculas, por ejemplo X).

Tipos de variables estadísticas

- **Variable cualitativa o atributos:** No se pueden cuantificar y se describen mediante palabras. Pueden ser:
 - Variables nominales: No existe una relación de orden entre las posibles categorías que toma la variable.
 - Variables ordinales: Existe una relación de orden entre las posibles categorías que toma la variable.
- **Variable cuantitativa:** Se pueden contar y medir; se describen utilizando números.
 - Variables discretas: Son las que toman un número finito de valores o infinito numerable.
 - Variables continuas: Pueden tomar un número infinito no numerable de valores.

1.2. Tabulación de datos

Definición

Una tabla estadística recoge la información sobre un carácter cualitativo o cuantitativo, denotado por X , de una población (o muestra), con objeto de resumirla.

Distribución de frecuencias para variables cualitativas nominales

Sea una población o una muestra de N elementos y sea X una variable estadística cuyas posibles categorías del carácter cualitativo o atributo son x_1, x_2, \dots, x_r . Se definen:

- Frecuencia absoluta de x_i , denotada por n_i , número de veces que aparece la modalidad x_i en el total de los N elementos de la población o muestra.
- Frecuencia relativa de x_i , denotada por f_i es la proporción de individuos que presentan la modalidad x_i en el total de N elementos de la población o muestra; es decir $f_i = \frac{n_i}{N}$.

En una tabla como la siguiente se representan las modalidades o categorías junto con sus respectivas frecuencias absolutas y relativas, que se llama distribución de frecuencias:

x_i	n_i	f_i
x_1	n_1	f_1
x_2	n_2	f_2
\vdots	\vdots	\vdots
x_r	n_r	f_r
N		1



Distribución de frecuencias para variables cualitativas ordinales y variables cuantitativas

- **Distribuciones con datos no agrupados en intervalos:** Sea X la variable que representa un carácter ordinal o cuantitativo que toma pocos valores diferentes.

Sea una población o una muestra de N elementos y sean x_1, x_2, \dots, x_r los valores que toma X ordenados de mayor a menor. Se definen:

- Frecuencia absoluta de x_i , denotada por n_i , número de veces que aparece la modalidad x_i en el total de los N elementos.



- Frecuencia relativa de x_i , denotada por f_i , es la proporción de individuos que presentan la modalidad x_i del total; es decir $f_i = \frac{n_i}{N}$.
- Frecuencia absoluta acumulada de x_i , denotada por N_i , es el número de observaciones menores o iguales que x_i , de modo que $N_i = n_1 + \dots + n_i = N_{i-1} + n_i$.
- Frecuencia relativa acumulada de x_i , denotada por F_i , es la proporción del total para los que la variable toma un valor menor o igual que x_i .

$$F_i = f_1 + \dots + f_i = F_{i-1} + f_i = \frac{N_i}{N}$$

En una tabla como la siguiente se representa el conjunto de valores junto con sus respectivas frecuencias, llamada distribución de frecuencias:

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots
x_r	n_r	f_r	N	$F_r = 1$
N		1		

- **Distribuciones con datos agrupados en intervalos:** Se emplean cuando X toma un número elevado de valores diferentes.

Sean N observaciones de la variable cuantitativa X que se agrupan en r intervalos, que por definición son intervalos abiertos a la izquierda y cerrados a la derecha, $(L_{i-1}, L_i]$. Se definen:

- Frecuencia absoluta del intervalo i -ésimo, $(L_{i-1}, L_i]$ es el número de observaciones pertenecientes a este intervalo.
- Frecuencia relativa del intervalo i -ésimo, $(L_{i-1}, L_i]$ es la proporción de individuos que toman valores en dicho intervalo. Se denota por f_i .
- Frecuencia absoluta acumulada del intervalo i -ésimo, $(L_{i-1}, L_i]$ es el número de observaciones menores o iguales que L_i ; se denota por N_i .
- Frecuencia relativa acumulada del intervalo i -ésimo, $(L_{i-1}, L_i]$ es la proporción de observaciones menores o iguales que L_i ; se denota por F_i .
- Número de intervalos, k , no se determina de forma fija sino que se toma de modo que se pueda trabajar cómodamente y ver bien la estructura de los datos. Se toma aproximadamente

$$k \approx \begin{cases} \sqrt{N} & \text{si } N \text{ no es muy grande} \\ 1 + 3,22 \ln(N) & \text{en otro caso} \end{cases}$$

Cuando los datos se agrupan en intervalos se tiene, además, que tener en cuenta:

- Amplitud del intervalo: $a_i = L_i - L_{i-1}$.
- Marca de clase: Punto medio del intervalo

$$x_i = \frac{L_{i-1} + L_i}{2}$$

- Densidad de frecuencia: Número de observaciones por unidad de amplitud del intervalo, $d_i = \frac{n_i}{a_i}$.

En una tabla como la siguiente se representa el conjunto de intervalos junto con sus respectivas frecuencias, llamada distribución de frecuencias:

$L_{i-1} - L_i$	x_i	n_i	f_i	N_i	F_i	a_i	d_i
$L_0 - L_1$	x_1	n_1	f_1	N_1	F_1	a_1	d_1
$L_1 - L_2$	x_2	n_2	f_2	N_2	F_2	a_2	d_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$L_{r-1} - L_r$	x_r	n_r	f_r	N_r	F_r	a_r	d_r
		N	1				

1.3. Representaciones gráficas

■ Representaciones gráficas de datos cualitativos:

1. Diagramas de rectángulos

Se representan las distintas modalidades en el eje de abscisas y sobre cada una de ellas se levanta un rectángulo cuya altura es igual a su frecuencia absoluta o relativa.

2. Diagramas de sectores

Se divide el círculo en sectores circulares tantos como modalidades haya, correspondiendo a cada una de ellas el sector con área proporcional a su frecuencia relativa o absoluta.

■ Representaciones gráficas de datos cuantitativos no agrupados en intervalos:

1. Diagrama de barras

Se obtiene dibujando una barra sobre cada valor x_i con altura igual a n_i o f_i .

2. **Poligonal de frecuencias**

Se obtiene uniendo con segmentos los puntos de coordenadas (x_i, n_i) o (x_i, f_i) .

■ **Representaciones gráficas para datos cuantitativos agrupados en intervalos:**

1. **Histograma**

Sobre cada intervalo se representa un rectángulo con altura igual a la densidad de frecuencia d_i , con objeto de que el área del rectángulo sea igual a la frecuencia absoluta del correspondiente intervalo. Cuando los intervalos tienen la misma amplitud se puede utilizar como altura del rectángulo la frecuencia absoluta n_i y se obtendrán áreas de rectángulos proporcionales a las frecuencias.

2. **Polígono de frecuencias**

Se obtiene al unir los puntos medios de las bases superiores de los rectángulos del histograma (x_i, d_i) y cerrar el polígono cortando el eje de abscisas de forma que el área encerrada entre el polígono de frecuencias y el eje horizontal coincida con el área del histograma.

1.4. Medidas descriptivas de posición y dispersión

Medidas descriptivas de posición

En esta sección se estudian medidas que indican cómo se distribuyen los valores que toma la variable X en la recta real (ordinales o cuantitativas).

Las principales medidas de posición central son: la media, la mediana y la moda, y los cuantiles son las medidas de posición no centrales que se estudian.

1. **Media aritmética:** Distinguimos entre datos no agrupados por intervalos y aquellos que lo están.

- Para una distribución de frecuencias con datos no agrupados por intervalos (x_i, n_i) ((x_i, f_i)) se define la media aritmética \bar{x} como:

$$\bar{x} = \frac{\sum_{i=1}^r n_i x_i}{N} = \sum_{i=1}^r f_i x_i$$

donde $N = \sum_{i=1}^r n_i$ ($\sum_{i=1}^r f_i = 1$).

- Cuando los datos están agrupados en intervalos $\{(L_{i-1}, L_i], n_i\}$, la expresión para calcular la media aritmética de la muestra es la misma que en el caso anterior tomando x_i la marca de la clase de intervalo $(L_{i-1}, L_i]$.

Importantes: La media aritmética se ve afectada por valores extremos.

Propiedades

- a) La media de la desviaciones de la muestra respecto de su media aritmética vale 0, esto es:

$$\sum_{i=1}^r \frac{(x_i - \bar{x})n_i}{N} = 0 \Leftrightarrow \sum_{i=1}^r (x_i - \bar{x})f_i = 0$$

- b) Le afectan los cambios de origen.
c) Le afectan los cambios de escala.
d) Si la muestra se divide en k subconjuntos disjuntos (llamados estratos), la media \bar{x} se puede calcular a partir de las medias de cada uno de los estratos, \bar{x}_i ,

$$\bar{x} = \frac{N_1\bar{x}_1 + \dots + N_k\bar{x}_k}{N_1 + \dots + N_k}$$

siendo N_i el número de observaciones del estrato i .

2. **Media geométrica** Si la variable estadística toma los valores x_1, x_2, \dots, x_r cuyas frecuencias absolutas son n_1, n_2, \dots, n_r , con $N = n_1 + \dots + n_r$, se define la media geométrica de X como:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdots x_r^{n_r}}$$

Se utiliza para obtener tasas de variación (tema 7).

3. **Mediana:** Se denota con Me y es el valor de la variable estadística que una vez ordenados los datos de menor a mayor, deja a su izquierda el 50 % de las observaciones.

- **Cálculo de la mediana para datos no agrupados por intervalos:** Se determina el primer valor x_i de la muestra cuya frecuencia absoluta acumulada es mayor o igual que $N/2$. Si esta frecuencia absoluta es mayor que $N/2$, la mediana es x_i . Si es igual a $N/2$ la mediana se define como la media aritmética de x_i y x_{i+1} , siempre que el valor resultante sea un valor factible de la variable. En caso contrario son medianas ambos valores.
- **Cálculo de la mediana para datos agrupados por intervalos:** Se determina el primer intervalo $(L_{i-1}, L_i]$ cuya frecuencia absoluta acumulada es mayor o igual que $N/2$ y el valor de la mediana se calcula:

$$Me = L_{i-1} + \frac{N/2 - N_{i-1}}{n_i}a_i$$

suponiendo que los datos se distribuyen uniformemente en el intervalo.

Importante: La mediana no se ve afectada por datos extremos.

Propiedades:

- a) Le afectan los cambios de origen.
- b) Le afectan los cambios de escala.

4. **Moda:** La moda Mo es el valor de la variable que se presenta con mayor frecuencia. A diferencia de las otras medidas de posición, la moda también se puede calcular para variables cualitativas.

- Cálculo de la moda para datos no agrupados por intervalos: Es el valor o valores con mayor frecuencia.
- Cálculo de la moda para datos agrupados por intervalos: Se define como la marca de clase del intervalo con mayor densidad de frecuencia.

Importante: La moda no se ve afectada por datos extremos.

Propiedades:

- a) Le afectan los cambios de origen.
- b) Le afectan los cambios de escala.

5. **Medidas descriptivas de posición no centrales: Cuantiles (C_q).**

Son valores que ordenados los datos de menor a mayor, dividen la distribución en intervalos con el mismo número de observaciones.

- **Cuartiles:** Son tres valores Q_1 , Q_2 y Q_3 que dividen la distribución en cuatro intervalos, cada uno de ellos con un 25 % de observaciones.
- **Deciles:** Son nueve valores D_1, D_2, \dots, D_9 que dividen la distribución en diez intervalos, cada uno de ellos con el 10 % de las observaciones.
- **Percentiles:** Son 99 valores P_1, \dots, P_{99} que dividen la distribución en cien intervalos cada uno de ellos con el 1 % de observaciones.

Su cálculo es similar al cálculo de la mediana, sustituyendo $N/2$ por $qN/100$ en el caso del cuantil q , C_q .

Para una distribución agrupada en intervalos:

$$C_q = L_{i-1} + \frac{\frac{qN}{100} - N_{i-1}}{n_i} a_i$$

Propiedades:

- a) Le afectan los cambios de origen.
- b) Le afectan los cambios de escala.

Medidas descriptivas de dispersión

Las medidas de dispersión se utilizan para describir la variabilidad de los datos de la muestra respecto de alguna medida de posición central, informando de su representatividad.

- **Recorrido o rango:** Amplitud del intervalo de la recta real más pequeño que contiene todas las observaciones.

Para datos no agrupados, el recorrido es la diferencia entre el valor máximo y mínimo tomado.

$$Re = \max_i \{x_i\} - \min_i \{x_i\}$$

Para datos agrupados por intervalos se define el recorrido como la diferencia entre el límite superior del último intervalo y el límite inferior del primero de los intervalos.

A mayor recorrido más dispersos estarán los datos.

- **Recorrido intercuartílico,** R_i Diferencia entre el cuartil tercero y el cuartil primero, es decir,

$$R_i = Q_3 - Q_1$$

Mide la variabilidad de los datos en torno a la mediana.

- **Varianza:** Medida de dispersión de los valores de la variable respecto a su media. Se define como:

$$S^2 = \frac{\sum_{i=1}^r (x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^r x_i^2 n_i}{N} - \bar{x}^2 = \sum_{i=1}^r x_i^2 f_i - \bar{x}^2$$

La varianza viene expresada en las unidades de la magnitud estudiada al cuadrado. Cuanto menor sea la varianza, menor dispersión de los datos en torno a la media y mayor representatividad de la media.

Propiedades:

1. La varianza siempre es mayor o igual que cero.
2. Los cambios de origen no afectan a la varianza.
3. La varianza se ve afectada por los cambios de escala.

- **Desviación típica:** Se define como

$$S = +\sqrt{\frac{\sum_{i=1}^r (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{\sum_{i=1}^r x_i^2 n_i}{N} - \bar{x}^2} = \sqrt{\sum_{i=1}^r x_i^2 f_i - \bar{x}^2}$$

Se expresa en la misma unidad que la variable que se estudia.

Propiedades:

1. La desviación típica siempre es mayor o igual que cero.
 2. No le afectan cambios de origen.
 3. Le afectan los cambios de escala
- **Coefficiente de variación de Pearson:** Es una medida de dispersión que no depende de la escala y que permite comparar las dispersiones relativas de varias muestras. Se define como:

$$Cv(X) = \frac{S}{|\bar{x}|} \quad \text{Desviación típica}$$

Por supuesto, para que se pueda calcular esta medida es preciso que la media no sea cero.

Propiedades:

1. Es adimensional por lo que es útil para comparar distribuciones expresadas en distintas unidades.
2. Vale 0 cuando la desviación típica es cero. En este caso todos los valores de la distribución son iguales a la media.
3. Le afectan los cambios de origen.
4. No le afectan los cambios de escala.

1.5. Medidas descriptivas de concentración: índice de Gini y curva de Lorenz

Las medidas de concentración estudian el grado de desigualdad o de concentración en el reparto de los valores de una variable entre los miembros de una población o muestra. Por ejemplo, sirve para estudiar la distribución de salarios, rentas, beneficios,... entre individuos (trabajadores, familias, empresas,...).

Se utilizan cuando se está interesado en la mayor o menor equidad en la distribución del total de los valores observados de la variable X entre la población o muestra.

Hay dos situaciones extremas de la distribución una variable, que son:

- Mínima concentración o máxima igualdad: cuando todos los integrantes de la población o muestra tienen asignado el mismo valor de la variable.
- Máxima concentración o mínima igualdad: cuando un único elemento de la muestra tiene asignado un valor de la variable distinto de cero y los restantes elementos de la muestra tienen asignado un valor 0.

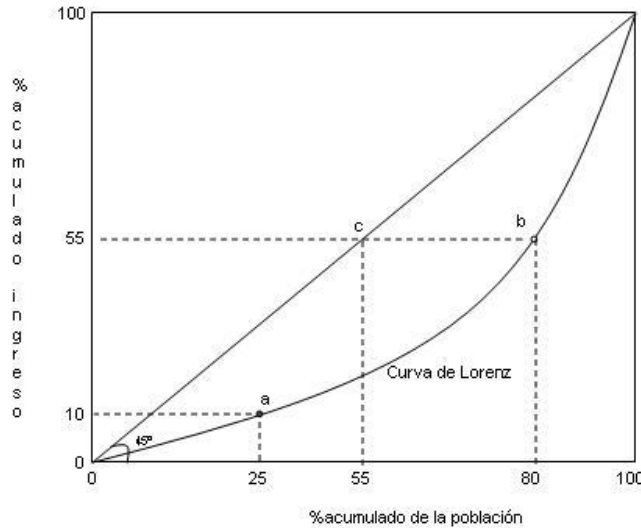
Índice de Gini

Para la obtener este índice de concentración de la distribución de los valores que toma una variable X se construye una tabla donde se incluyen:

1. En la primera columna se coloca en orden decreciente los posibles valores que toma la variable X .
2. A continuación columna de frecuencias absolutas, seguida de la columna de las frecuencias absolutas acumuladas.
3. Otra columnas donde se representan los recursos totales percibidos por cada n_i individuos que toman el valor de X denotado por x_i .
4. Columna de los recursos acumulados $u_k = \sum_{i=1}^k x_i n_i$.
5. Columna de porcentajes acumulados de individuos $p_i = 100(N_i/N)$ y columna de recursos acumulados $q_i = 100(u_i/u_r)$

Orden	x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{100N_i}{N}$	$q_i = \frac{100u_i}{u_r}$
1.º	x_1	n_1	$x_1 n_1$	N_1	u_1	p_1	q_1
2.º	x_2	n_2	$x_2 n_2$	N_2	u_2	p_2	q_2
...
i .º	x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
...
r .º	x_r	n_r	$x_r n_r$	N	u_r	$p_r = 100$	$q_r = 100$

Si el reparto fuese igualitario, todos los perceptores recibirían la misma cantidad y se cumpliría $p_i = q_i$ para $i = 1, 2, \dots, r - 1$, mientras que si hay un único perceptor $q_i = 0$ para $i = 1, 2, \dots, r - 1$; en cualquier otro caso intermedio, si estos valores están



próximos la concentración será muy baja y si están alejados la concentración será elevada. Es decir, si las diferencias $p_i - q_i$ son grandes habrá una mayor concentración del total de la variable X . El índice de concentración de Gini se define como:

$$I_G = \frac{\sum_{i=1}^{r-1} (p_i - q_i)}{\sum_{i=1}^{r-1} p_i} = 1 - \frac{\sum_{i=1}^{r-1} q_i}{\sum_{i=1}^{r-1} p_i}$$

donde $0 \leq I_G \leq 1$.

Curva de Lorenz

La curva de Lorenz o curva de concentración es una gráfica que se deduce a partir de la información suministrada para el cálculo del índice de Gini, reflejando la mayor o menor concentración en la distribución de una magnitud.

Para su representación gráfica, en el eje de abscisas las proporciones acumuladas de los perceptores o elementos de la muestra, p_i , y en el eje de ordenadas se sitúan las proporciones acumuladas de los valores que toma la variable estadística, q_i , y como se cumple que $p_i \geq q_i$, se tiene que si trazamos la diagonal del cuadrado de vértices

$$(0, 0), (100, 0), (100, 100), (0, 100)$$

los puntos de coordenadas (p_i, q_i) se sitúan por debajo de esta diagonal. La curva de Lorenz resulta de unir todos los puntos (p_i, q_i) , (una línea quebrada).

Mientras que el índice de Gini da un valor indicativo del nivel de concentración producido en el reparto, la curva de Lorenz describe gráficamente el fenómeno

1. VARIABLES ESTADÍSTICAS UNIDIMENSIONALES

pudiendo identificar para que grupos de perceptores se acentúa la concentración y para cuáles es menor.

1.6. Ejemplos

Ejemplo 1

La situación laboral de 3160 hombres mayores de edad inscritos en un municipio es la siguiente: Ocupados 1580, Estudiantes 790, Parados 158 y Jubilados 632. Construir la tabla de frecuencias correspondiente a esta variable.

	n_i	f_i
Estudiantes	790	0,25
jubilados	632	0,20
ocupados	1580	0,50
parados	158	0,05

Ejemplo 2

Por ejemplo supongamos que a una población de 40 estudiantes de esta facultad le preguntamos por el número de miembros de su familia, se tendría la variable estadística: Número de miembros de la familia cuyos valores son:

3	2	3	3	3	3	4	4	3	3
3	4	3	2	2	4	3	5	3	5
2	4	5	4	3	4	3	2	2	4
4	5	4	3	2	2	5	4	3	2

- ¿Cuál es el número de miembros más frecuente en una familia?
- ¿Cuántos alumnos tienen una familia formada como máximo por 3 miembros?
- Porcentaje de familias formadas por dos personas?

Solución

Tabulación de los datos

x_i	n_i	f_i
2	9	0,225
3	15	0,375
4	11	0,275
5	5	0,125

- a) El número más frecuente de miembros de una familia es 3.
- b) 24 familias tienen como máximo 3 miembros en su familia.
- c) El porcentaje de familias formadas por dos personas es 22,5 %.

Ejemplo 3

Tabular los siguientes datos correspondientes a las alturas (en cm) de 50 trabajadores de la empresa TEXTIL S.A. Determinar cuántos miden más de 1,80 m, indicando el correspondiente porcentaje

174	185	166	176	145	166	191	175	158	156
156	187	162	172	197	181	151	161	183	172
162	147	178	176	142	170	171	158	184	173
169	162	172	181	187	177	164	171	193	183
173	179	188	179	167	178	180	168	148	173

Solución

Tabulando los datos se tiene que 12 trabajadores miden más de 1,80 m.

Ejemplo 4

Una entidad bancaria dispone de 50 sucursales en el territorio nacional y ha observado el número de empleados que hay en cada una de ellas para un estudio posterior. Las observaciones obtenidas son:

12	10	9	11	15	16	9	10	10	11	12	13	14
15	11	11	12	16	17	17	16	16	15	14	12	11
11	11	12	12	12	15	13	14	16	15	18	19	18
10	11	12	12	11	13	13	15	13	11	12		

- a) Calcular la distribución de frecuencias absolutas, relativas y sus correspondientes acumuladas.
- b) ¿Qué proporción de sucursales tiene más de 15 empleados?
- c) Dibujar el diagrama de barras y el diagrama de frecuencias acumuladas.

- d) Agrupar en intervalos de amplitud 3 los valores de la variable, calcular su distribución de frecuencias y representar su histograma y su polígono de frecuencias acumuladas.

Solución

- a) Tabulación de los datos que se ve en la tabla correspondiente.

x_i	n_i	f_i	Ni	F_i
9	2	0,04	2	0,04
10	4	0,08	6	0,12
11	10	0,20	16	0,32
12	10	0,20	26	0,52
13	5	0,10	31	0,62
14	3	0,06	34	0,68
15	6	0,12	40	0,80
16	5	0,10	45	0,90
17	2	0,04	47	0,94
18	2	0,04	49	0,98
19	1	0,02	50	1

- b) El 20 % de las sucursales tienen más de 15 empleados.

- c)

$L_{i-1} - L_i$	x_i	n_i	Ni	F_i
[9, 12]	10,5	26	26	0,52
(12, 15]	13,5	14	40	0,8
(15, 18]	16,5	9	49	0,98
(18, 21]	19,5	1	50	1

- d)

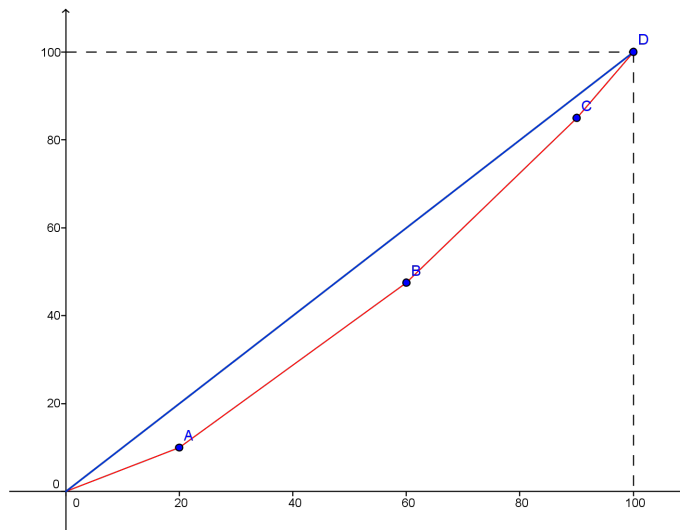
Ejemplo 5

Calcular e interpretar el índice de Gini asociado a la siguiente distribución de salarios mensuales de una empresa de 50 empleados, donde el salario se mide en cientos de euros. Representar la curva de Lorenz asociada.

1. VARIABLES ESTADÍSTICAS UNIDIMENSIONALES

Salario x_i	N.º de empleados n_i
8	10
15	20
20	15
24	5

x_i	n_i	$x_i n_i$	N_i	u_i	p_i	q_i
8	10	80	10	80	20	10
15	20	300	30	380	60	47,5
20	15	300	45	680	90	85
24	5	120	50	800	100	100



Solución

$$IG = 1 - \frac{142,5}{170}$$