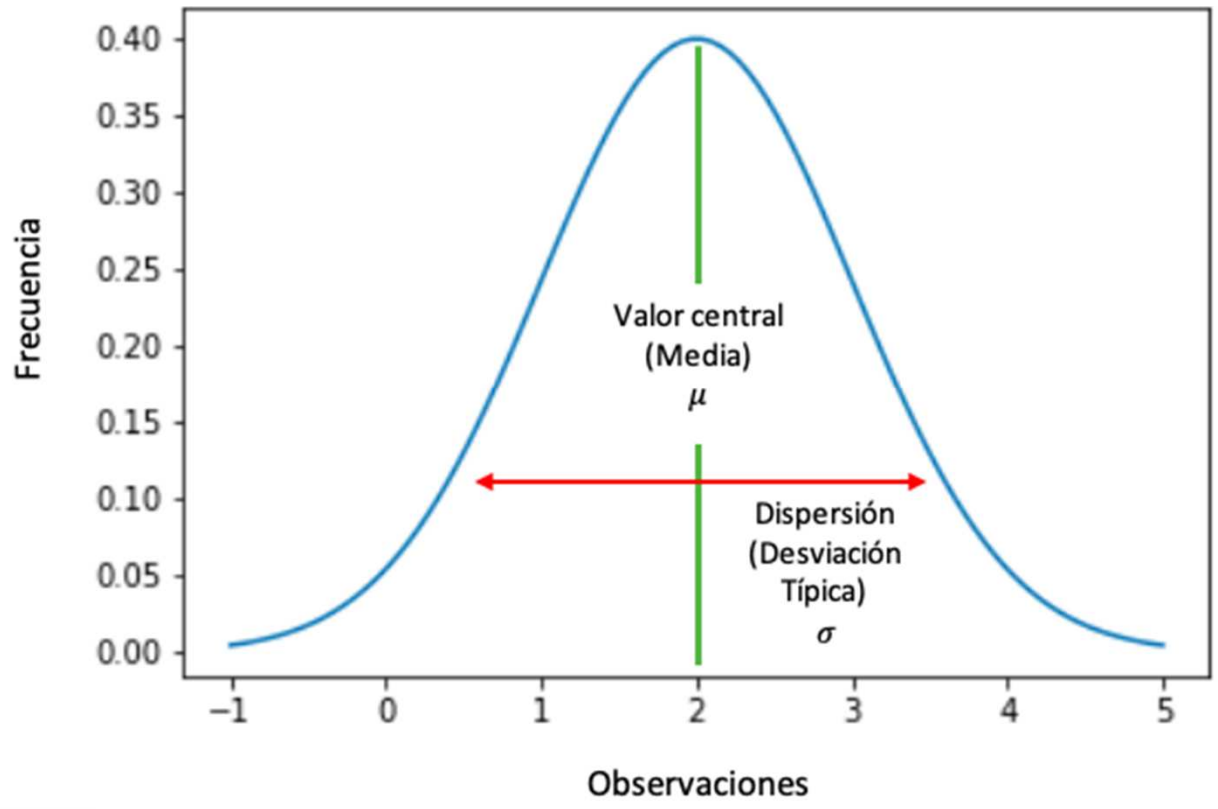
Introducción a la Estadística
Grado en Turismo

TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

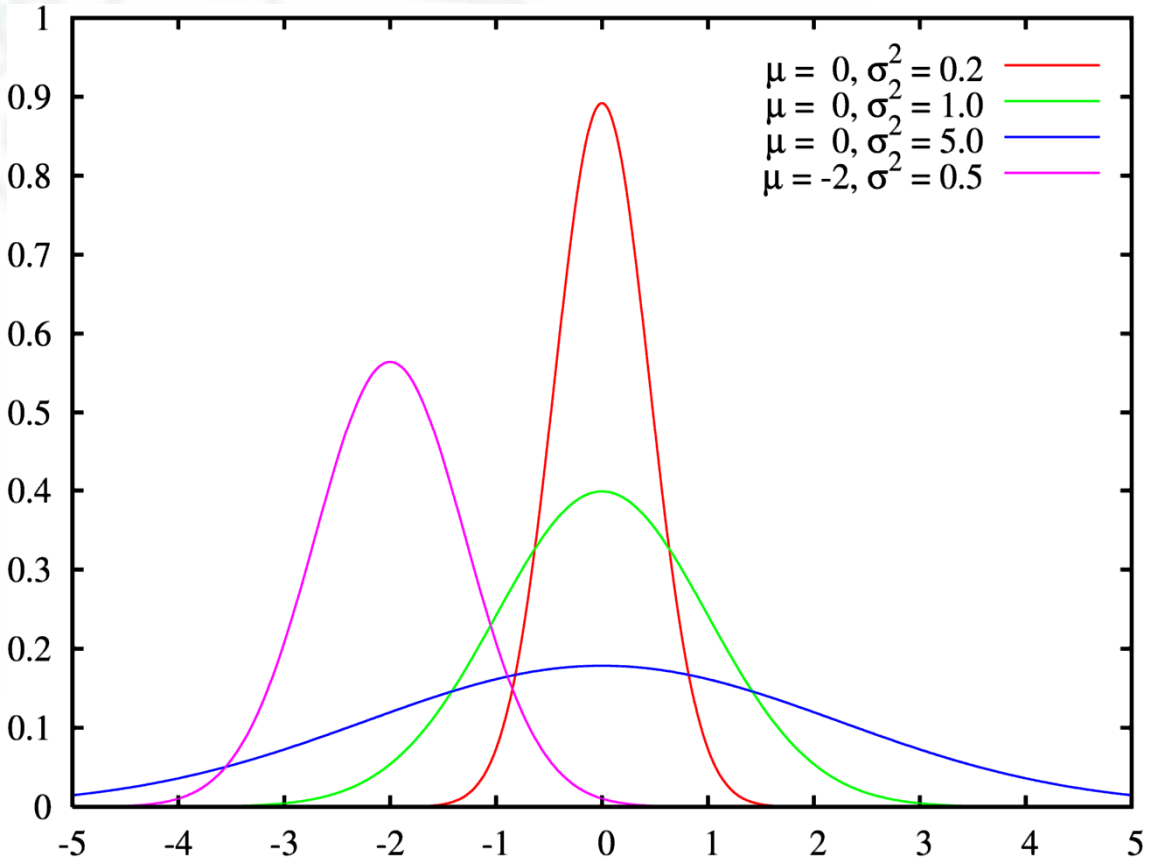
TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

EL PROBLEMA DEL INVESTIGADOR ESTADÍSTICO



TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

EL PROBLEMA DEL INVESTIGADOR ESTADÍSTICO



TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

MEDIDAS DE POSICIÓN CENTRAL

- Media (aritmética, geométrica, armónica)
- Mediana
- Moda

Introducción a la Estadística
Grado en Turismo

MEDIA ARITMÉTICA, GEOMÉTRICA Y ARMÓNICA

TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

MEDIA ARITMÉTICA

Su fórmula de cálculo es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N}$$

TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

VEAMOS UN EJEMPLO

PROPIEDADES DE LA MEDIA ARITMÉTICA

- La suma de las desviaciones de los valores de la variable con respecto a la media aritmética es igual a cero.

- Si a todos los valores de la variable se les suma o resta una misma cantidad, la media aritmética queda aumentada o disminuida también en dicha cantidad (*cambio de origen*).

PROPIEDADES DE LA MEDIA ARITMÉTICA

- Si todos los valores de la variable se multiplican o dividen por una misma constante, la media aritmética queda multiplicada o dividida también por dicha constante (*cambio de escala*).

- Si una variable Y es transformación lineal de otra variable X, tal que *Y=a+bX*, entonces la media aritmética de la variable Y sigue la misma transformación lineal con respecto a la media aritmética de la variable X, tal que:

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum (a + bx_i) n_i}{N} = \frac{\sum (an_i + bx_i n_i)}{N} = \frac{a\sum n_i}{N} + \frac{b\sum x_i n_i}{N} = a + b\bar{x}$$

VENTAJAS DE LA MEDIA ARITMÉTICA

- Su cálculo es muy sencillo (aunque solo se puede calcular en las variables de naturaleza cuantitativa).
- Para su cálculo se utilizan todos los valores de la distribución.
- Está perfectamente definida de forma objetiva y es única para cada distribución de frecuencias.
- Representa el centro de gravedad de la distribución.

DESVENTAJAS DE LA MEDIA ARITMÉTICA

- Es muy sensible a los valores extremos de la distribución con lo que puede llegar a ser poco representativa del conjunto si la dispersión de los datos es muy elevada.

- No puede ser calculada cuando la variable es de tipo cualitativo.

- Su cálculo no es posible en distribuciones agrupadas con intervalos abiertos.

MEDIA GEOMÉTRICA

Su fórmula de cálculo es la siguiente:

$$G = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}} = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}} = \left(x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k} \right)^{\frac{1}{N}}$$

O alternativamente:

$$\log G = \frac{1}{N} \log \left(x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k} \right) = \frac{1}{N} \left(n_1 \log x_1 + n_2 \log x_2 + \ldots + n_k \log x_k \right)$$

$$= \frac{\sum_{i=1}^{k} n_i \log x_i}{N} \quad \Rightarrow \quad G = anti\log \frac{\sum_{i=1}^{k} n_i \log x_i}{N}$$

TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

VEAMOS UN EJEMPLO

VENTAJAS DE LA MEDIA GEOMÉTRICA

- Si su cálculo es posible, está definida de forma objetiva y es única.
- Tiene en cuenta en su cálculo a todos los valores de la distribución.
- Los valores extremos tienen menor influencia que en la media aritmética por estar definida a través de productos en vez de sumas.
- Es más representativa que la media aritmética cuando la variable evoluciona de forma acumulativa con efectos multiplicativos.

DESVENTAJAS DE LA MEDIA GEOMÉTRICA

- Su cálculo es más complicado que el de la media aritmética.
- Si algún valor de la variable es igual a cero, el resultado obtenido no es representativo al obtenerse una media geométrica nula.
- Asimismo, si la variable presentara valores negativos podría darse el caso de que no fuera posible calcularla ya que se obtendrían soluciones imaginarias.

MEDIA ARMÓNICA

Su fórmula de cálculo es la siguiente:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}}$$

TEMA 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN.

VEAMOS UN EJEMPLO

VENTAJAS DE LA MEDIA ARMÓNICA

- Está definida de forma objetiva y es única.
- Intervienen todos los valores de la distribución.
- Es más representativa que las otras medias en los casos de obtener promedios en velocidades, rendimientos y productividades.

DESVENTAJAS DE LA MEDIA ARMÓNICA

- Si algún valor de la variable es nulo, no es posible calcular la media armónica.

- La presencia valores de la variable muy pequeños pueden provocar que sus inversos aumenten muchísimo, haciendo despreciable frente a ellos la información de otros valores de x_i

Introducción a la Estadística
Grado en Turismo

MEDIANA, MODA Y CUANTILES

Mediana

- La Mediana (Me) es una medida de posición central en cuyo cálculo no intervienen todos los valores de la variable y en la que se equilibra las frecuencias observadas a ambos lados de su valor.

- Dicho de otro modo, la Mediana es aquel valor tal que, tras ordenar los valores de la variable en orden creciente, deja a su izquierda y a su derecha el mismo número de frecuencias.

- Dependiendo del tipo de distribución analizada, la forma de calcularla varía.

Mediana

1. Distribución de frecuencias unitaria con valores no agrupados en intervalos:

- Si N es impar → Valor central de la distribución
- Si N es par → Media de los dos valores centrales

Mediana

2. Distribución de frecuencias no unitaria con valores no agrupados en intervalos:

Calculamos N/2 y las frecuencias absolutas acumuladas y comparamos:

- Si N/2 coincide con algún N_i → La Mediana es el promedio entre el valor asociado a ese N_i y el siguiente valor de la variable.
- Si N/2 no coincide → La Mediana es aquel valor de la variable cuyo N_i supera a N/2.

Mediana

2. Distribución de frecuencias no unitaria con valores no agrupados en intervalos:

Calculamos N/2 y las frecuencias absolutas acumuladas y comparamos:

- Si N/2 coincide con algún N_i → La Mediana es el promedio entre el valor asociado a ese N_i y el siguiente valor de la variable.
- Si N/2 no coincide → La Mediana es aquel valor de la variable cuyo N_i supera a N/2.

Mediana

3. Distribución de frecuencias con valores agrupados en intervalos:

- Si N/2 coincide con algún N_i → La Mediana es el extremo superior del intervalo que verifica la condición.
- Si N/2 no coincide → La Mediana se obtiene aplicando la siguiente fórmula que permite prorratear los valores dentro del intervalo:

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

Mediana

3. Distribución de frecuencias con valores agrupados en intervalos:

En la fórmula anterior:

- L_i es el extremo inferior del intervalo.
- N_{i-1} es la frecuencia absoluta acumulada hasta el intervalo anterior.
- a_i es la amplitud del intervalo ($L_{i+1} - L_i$) que cuyo N_i supera a N/2.

Mediana

Ventajas y desventajas de la Mediana

Ventajas

- Medida más representativa en el caso de variables que solo admiten la escala ordinal.
- Interpretación y cálculo sencillos.
- No es sensible a valores extremos de la variable.

Desventajas

- En su determinación no intervienen todos los valores de la variable, por lo que no se utiliza toda la información presente en la distribución.

Mediana

VEAMOS UN EJEMPLO

Moda

- La Moda (Mo) se define como el valor de la variable que más veces se repite, es decir, se trata de aquel valor de la variable que presenta la mayor frecuencia absoluta.

- Dada su definición, no tiene sentido hablar de moda en las distribuciones de frecuencias de tipo unitario.

- Dependiendo del tipo de distribución analizada, la forma de calcularla también varía.

Moda

1. Distribución de frecuencias con valores no agrupados en intervalos:

   - Como regla general, la Moda será simplemente aquel valor de la variable que presente la máxima frecuencia absoluta.

   - Si hubiera más de un valor con la misma frecuencia absoluta, diríamos que se trata de una distribución bimodal (2 valores), trimodal (3 valores), etc.

Moda

2. Distribución de frecuencias con valores agrupados en intervalos:

- En este caso, hablaremos de intervalo modal → Será el intervalo con mayor frecuencia absoluta siempre que todos los intervalos tengan todos la misma amplitud.
- No obstante, es posible obtener un valor preciso para la Moda aplicando la siguiente fórmula:

$$Mo = L_i + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i$$

Moda

2. Distribución de frecuencias con valores agrupados en intervalos:

En la fórmula anterior:

- L_i es el extremo inferior del intervalo.
- n_{i-1} es la frecuencia absoluta del intervalo anterior.
- n_{i+1} es la frecuencia absoluta del intervalo posterior.
- a_i es la amplitud del intervalo ($L_{i+1} - L_i$) en el que se encuentra la Moda.

Moda

2. Distribución de frecuencias con valores agrupados en intervalos:

- Si los intervalos fueran de diferente amplitud, la fórmula pasa a ser:

$$Mo = L_i + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a_i$$

Donde di hace referencia a la densidad de frecuencia definida como $d_i = n_i / a_i$.

Moda

Ventajas y desventajas de la Moda

Ventajas

- Sencillez de cálculo y fácil interpretación.
- Es posible calcularla tanto para variables cuantitativas como para variables cualitativas.

Desventajas

- En su determinación no intervienen todos los valores de la distribución, centrándonos sólo en la mayor frecuencia absoluta.

Moda

VEAMOS UN EJEMPLO

Cuantiles

Los cuantiles son aquellos valores que dividen la distribución en un cierto número de partes iguales, de manera que en cada una de ellas hay el mismo porcentaje de valores de la variable. Los más importantes son los siguientes:

- Cuartiles (C_i): 3 valores que dividen la distribución en 4 partes iguales, por lo que dentro de cada uno está incluido el 25% de los valores (25%-50%-75%).
- Deciles (D_i): 9 valores que dividen la distribución en 10 partes iguales, por lo que dentro de cada uno está incluido el 10% de los valores (10%, 20%, ..., 90%).
- Percentiles (P_i): 99 valores que dividen la distribución en 100 partes iguales, por lo que dentro de cada uno está incluido el 1% de los valores (1%, 2%, 3%,..., 99%).

Cuantiles

1. Cálculo en Distribuciones No Agrupadas en Intervalos

- Calculamos $r \cdot N/q$ (r = nº cuantil, q = tipo cuantil) y comparamos resultado con la columna de N_i.

- Si $r \cdot N/q = N_i \rightarrow$ El cuantil se obtiene como $(x_i + x_{i+1})/2$

- En caso contrario, buscamos primer $N_i > r \cdot N/q \rightarrow$ El cuantil será el valor de x_i asociado a ese N_i.

Cuantiles

2. Cálculo en Distribuciones Agrupadas en Intervalos

Primero determinamos donde está el cuantil y aplicamos la siguiente fórmula (similar al cálculo de la Mediana):

$$Q = L_i + \frac{\frac{rN}{q} - N_{i-1}}{n_i} \cdot a_i$$

Cuantiles

2. Cálculo en Distribuciones Agrupadas en Intervalos

En la fórmula anterior:

- L_i es el extremo inferior del intervalo.
- N_{i-1} es la frecuencia absoluta acumulada hasta el intervalo anterior.
- a_i es la amplitud del intervalo $(L_{i+1} - L_i)$ que cuyo N_i supera a $r \cdot N/q$.

Cuantiles

VEAMOS UN EJEMPLO

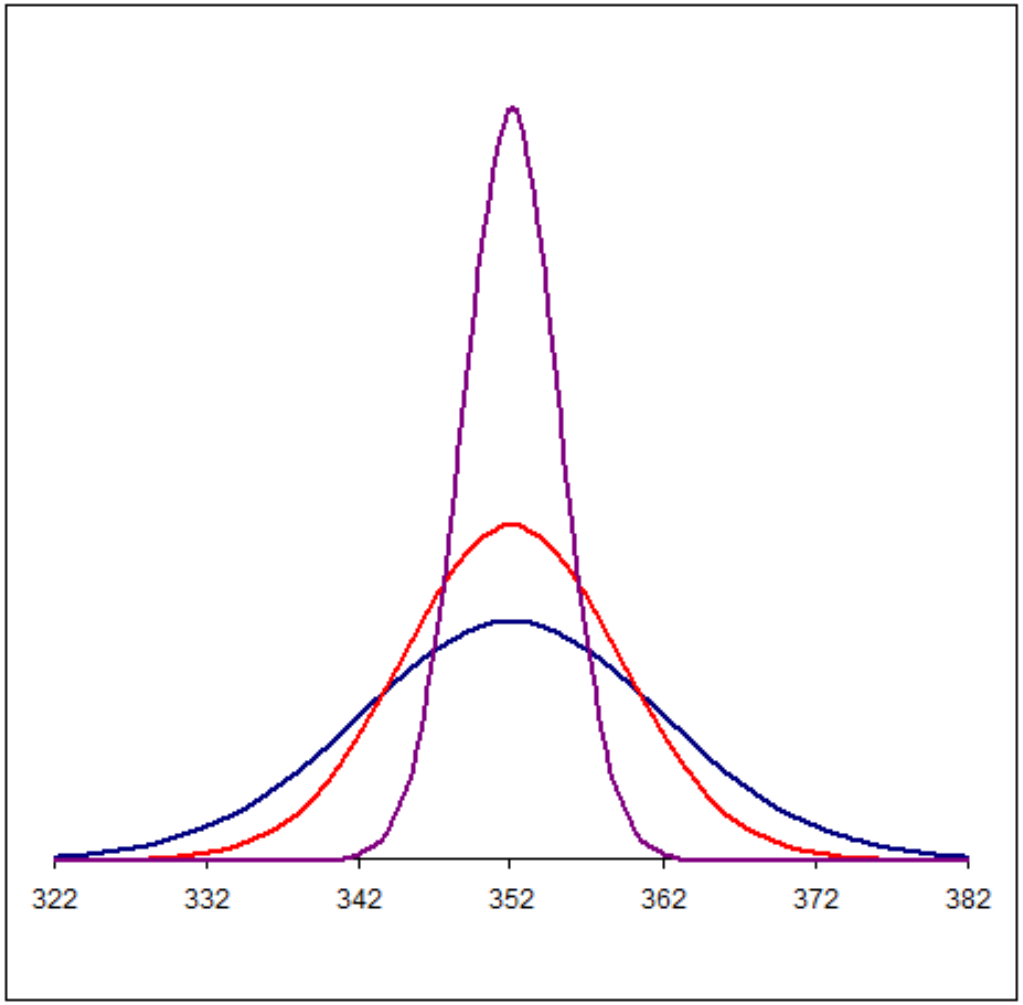
Introducción a la Estadística
Grado en Turismo

MEDIDAS DE DISPERSIÓN

Medidas de Dispersión

REFLEXIÓN: ¿CÓMO DE REPRESENTATIVA ES LA MEDIA DE UNA DISTRIBUCIÓN?

Medidas de Dispersión



Medidas de Dispersión

- Medidas de Dispersión Absoluta: Rango, Recorrido Intercuartílico, Desviación Absoluta, **Varianza**, **Desviación Típica**

- Medidas de Dispersión Relativa: **Coeficiente de Variación de Pearson**

Medidas de Dispersión

Medidas de Dispersión Absoluta

- Rango: $R = x_k - x_1$
- Recorrido Intercuartílico: $RI = C_3 - C_1$
- Desviación Absoluta: $D_{\bar{x}} = \dfrac{\sum_{i=1}^{k} \left| x_i - \bar{x} \right| n_i}{N}$

Medidas de Dispersión

Medidas de Dispersión Absoluta

- Varianza

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

$$\frac{\sum_{i=1}^{k} x_i^2 n_i}{N} + \bar{x}^2 \cdot \frac{N}{N} - 2\bar{x}\cdot\bar{x} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} + \bar{x}^2 - 2\bar{x}^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2$$

Medidas de Dispersión

Medidas de Dispersión Absoluta

- Desviación Típica

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

Medidas de Dispersión

VEAMOS ALGUNOS EJEMPLOS

Coeficiente de Variación

Medidas de Dispersión Relativa

Coeficiente de Variación de Pearson

$$\gamma = \frac{S}{|\overline{X}|}$$

Coeficiente de Variación

- Al estar expresado en porcentaje, permite comparar la dispersión de varias distribuciones aunque las variables estén en diferentes unidades de medida.

- El valor mínimo del coeficiente es cero (S=0 → No hay dispersión → Es una constante).

- La dispersión es óptima si es igual o menor al 30%. Si es superior al 50%, podemos considerar que la media es muy poco representativa.

- Si la media fuera nula → Cambio de origen de la variable.

Coeficiente de Variación

VEAMOS UN EJEMPLO

Tipificación de una Variable

La tipificación es el procedimiento mediante el cual podemos transformar cualquier variable en una nueva que denominaremos Z con media igual a cero y varianza igual a uno.

Para tipificar, restaremos a cada valor de la variable la media de la distribución, y dividiremos el resultado por su desviación típica tal que:

$$z_i = \frac{x_i - \bar{x}}{S}$$

Tipificación de una Variable

VEAMOS UN EJEMPLO

Introducción a la Estadística
Grado en Turismo

MEDIDAS DE FORMA Y DE CONCENTRACIÓN

Medidas de Forma

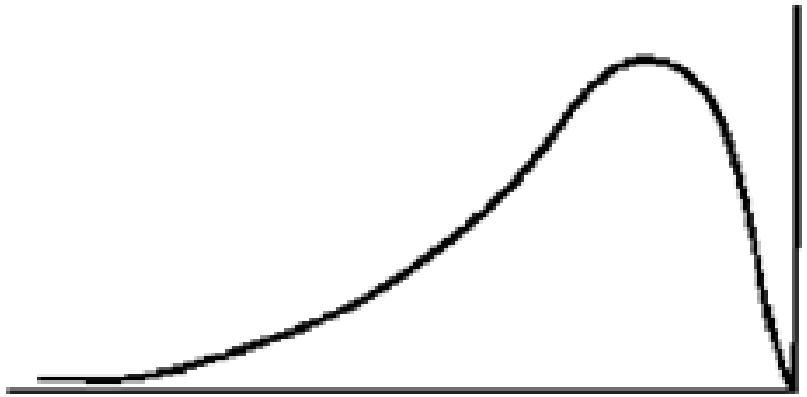
Aunque las medidas de posición y dispersión de dos distribuciones sean iguales, no tenemos datos analíticos para ver si son distintas por lo que debemos recurrir a medidas de forma.

- Asimetría: deformación horizontal de los valores de la variable respecto a un valor central (media).

- Apuntamiento o curtosis: concentración de las frecuencias en la zona central de la distribución comparada con la Normal

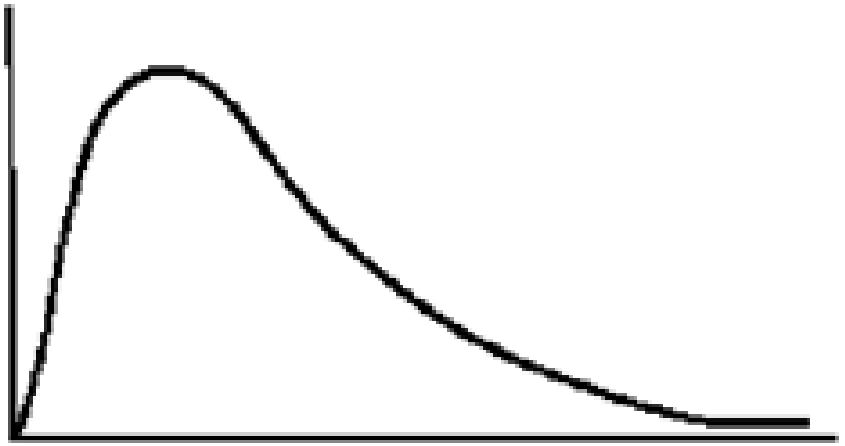
Asimetría

Asimétrica a la izquierda ($Mo \geq Me \geq \overline{X}$).



Asimetría

Asimétrica a la derecha ($Mo \leq Me \leq \overline{X}$).



Asimetría

1. Coeficiente de Asimetría de Pearson

$$A_P = \frac{\bar{x} - Mo}{S}$$

- Si $A_P>0$, la distribución es asimétrica a la derecha.
- Si $A_P=0$, la distribución es simétrica.
- Si $A_P<0$, la distribución es asimétrica a la izquierda.

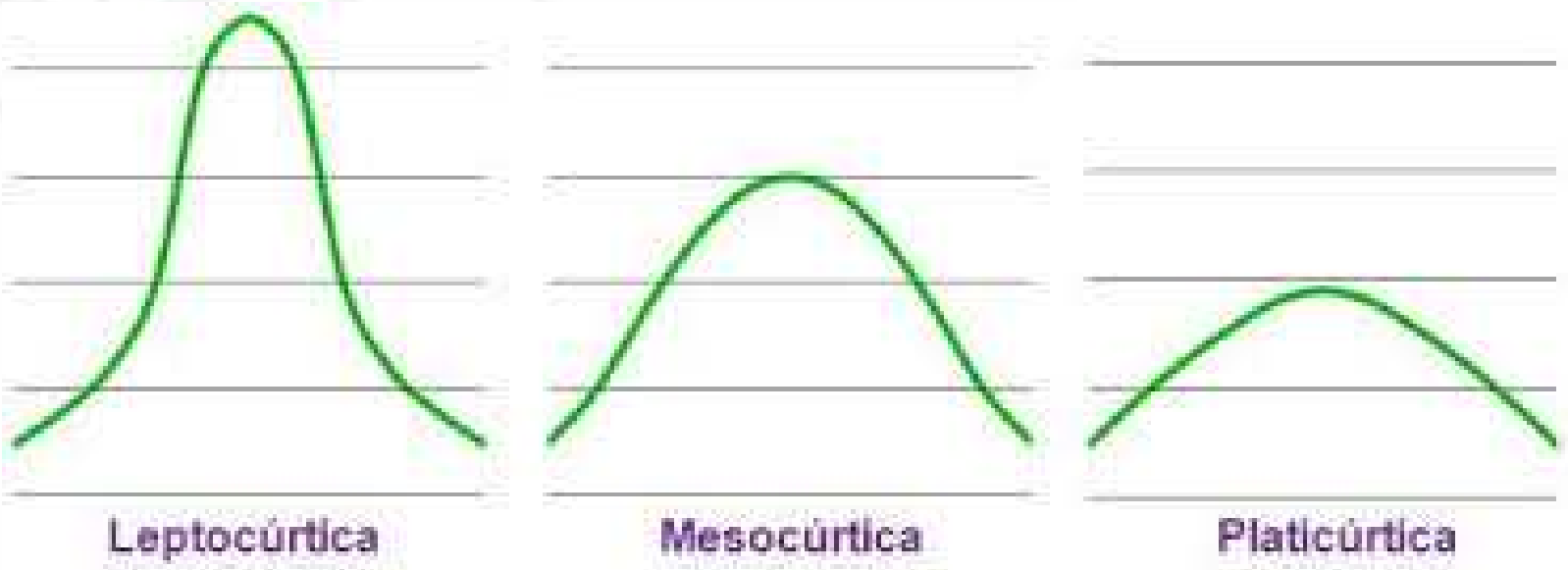
Asimetría

2. Coeficiente de Asimetría de Fisher

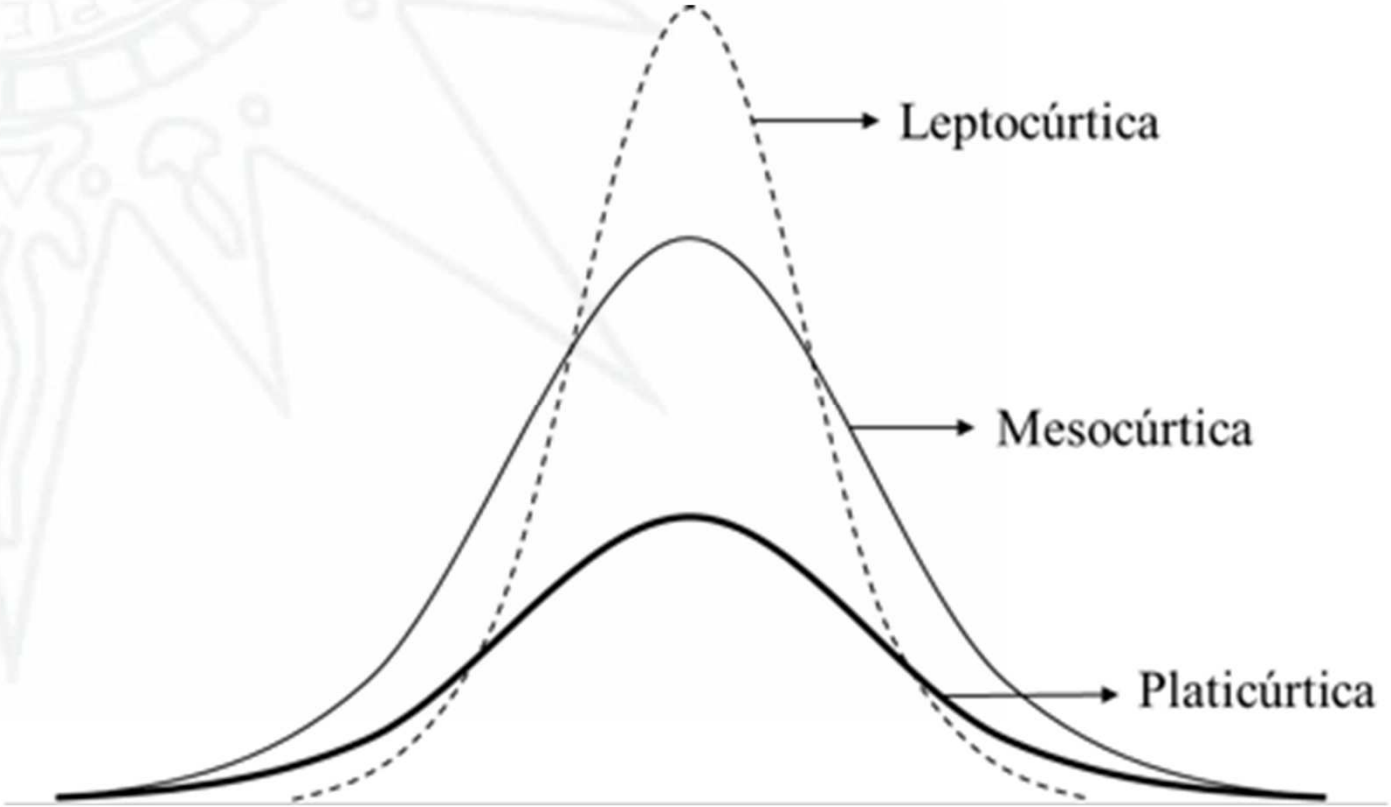
$$g_1 = \frac{\sum_{i=1}^{k} \left(x_i - \bar{x}\right)^3 n_i}{\frac{N}{S^3}}$$

- Si $g_1 > 0$, la distribución es asimétrica a la derecha.
- Si $g_1 = 0$, la distribución es simétrica.
- Si $g_1 < 0$, la distribución es asimétrica a la izquierda.

Apuntamiento o Curtosis



Apuntamiento o Curtosis



Apuntamiento o Curtosis

Coeficiente de Apuntamiento de Fisher

$$g_2 = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^4 n_i}{\frac{N}{S^4}} - 3$$

- Si $g_2>0$, la distribución es leptocúrtica.
- Si $g_2=0$, la distribución es mesocúrtica.
- Si $g_2<0$, la distribución es platicúrtica.

Medidas de Forma

VEAMOS ALGUNOS EJEMPLOS

Medidas de Concentración

Medidas de Concentración

Las medidas de concentración tratan de poner de manifiesto el mayor o menor grado de igualdad en el reparto total de los valores de la variable.

Son por tanto, indicadores del grado de equidistribución de la variable. De ahí que este tipo de medidas se apliquen sobre todo en Economía para distribuciones de rentas, salarios, etc.

Medidas de Concentración

Curva de Lorenz

Permite estudiar gráficamente la concentración de la distribución mediante la representación en el eje de abscisas del porcentaje de frecuencias acumuladas vs. los porcentajes acumulados del total de la variable en el eje de ordenadas.

Al unir los puntos resultantes obtenemos la curva, cuya forma nos permitirá determinar el nivel de concentración.

Medidas de Concentración

Curva de Lorenz

x_i	n_i	$x_i\, n_i$	N_i	u_i	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{u_i}{u_k} \cdot 100$	$p_i - q_i$
x_1	n_1	$x_1\, n_1$	N_1	u_1	p_1	q_1	$p_1 - q_1$
x_2	n_2	$x_2\, n_2$	N_2	u_2	p_2	q_2	$p_2 - q_2$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_k	n_k	$x_k\, n_k$	N_k	u_k	$p_k = 100$	$q_k = 100$	$p_k - q_k = 0$
	N	u_k					

Medidas de Concentración

Curva de Lorenz

Donde:

$$u_1 = x_1 n_1$$

$$u_2 = x_1 n_1 + x_2 n_2$$

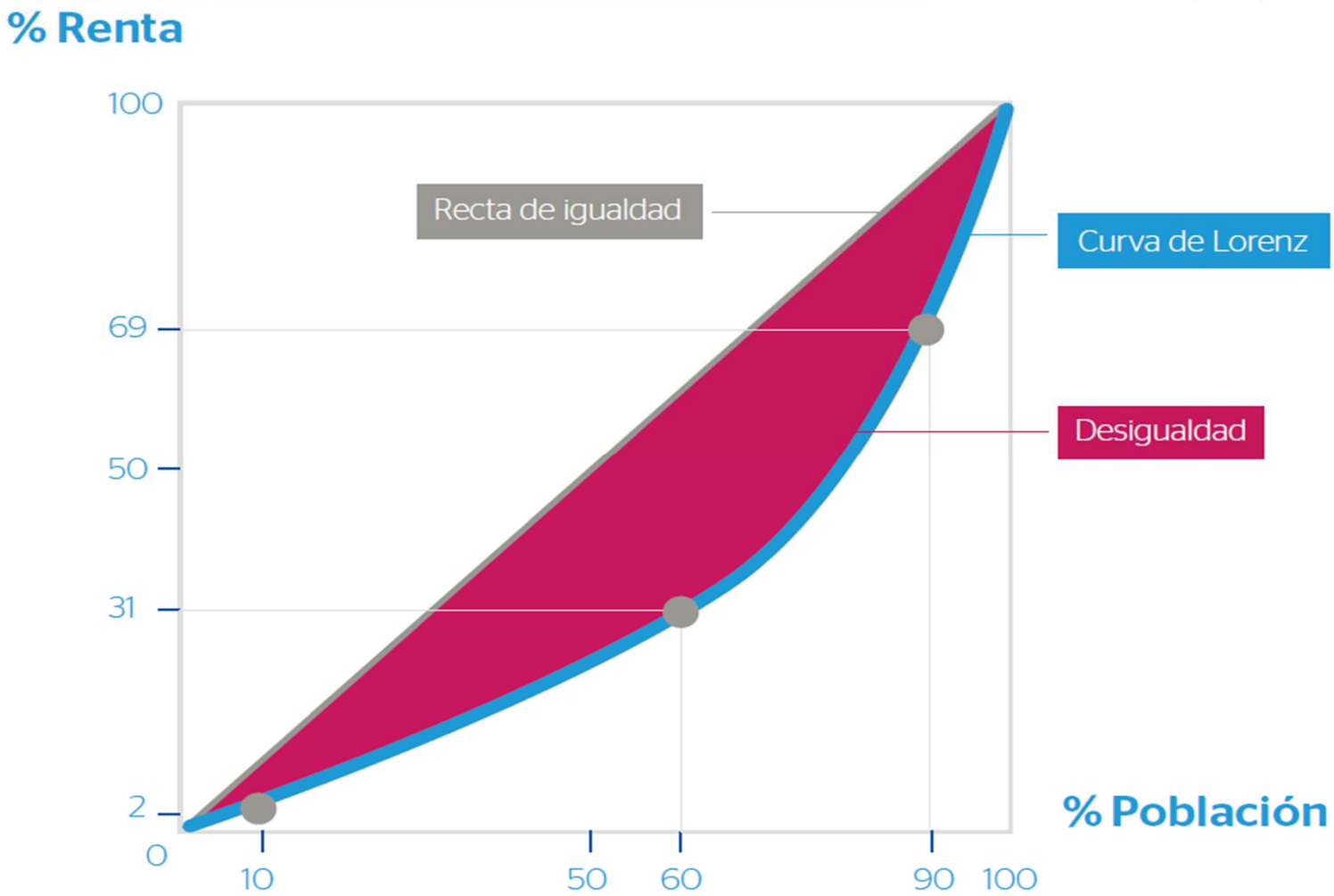
$$\vdots$$

$$u_k = x_1 n_1 + x_2 n_2 + \ldots + x_k n_k = \sum_{i=1}^{k} x_i n_i$$

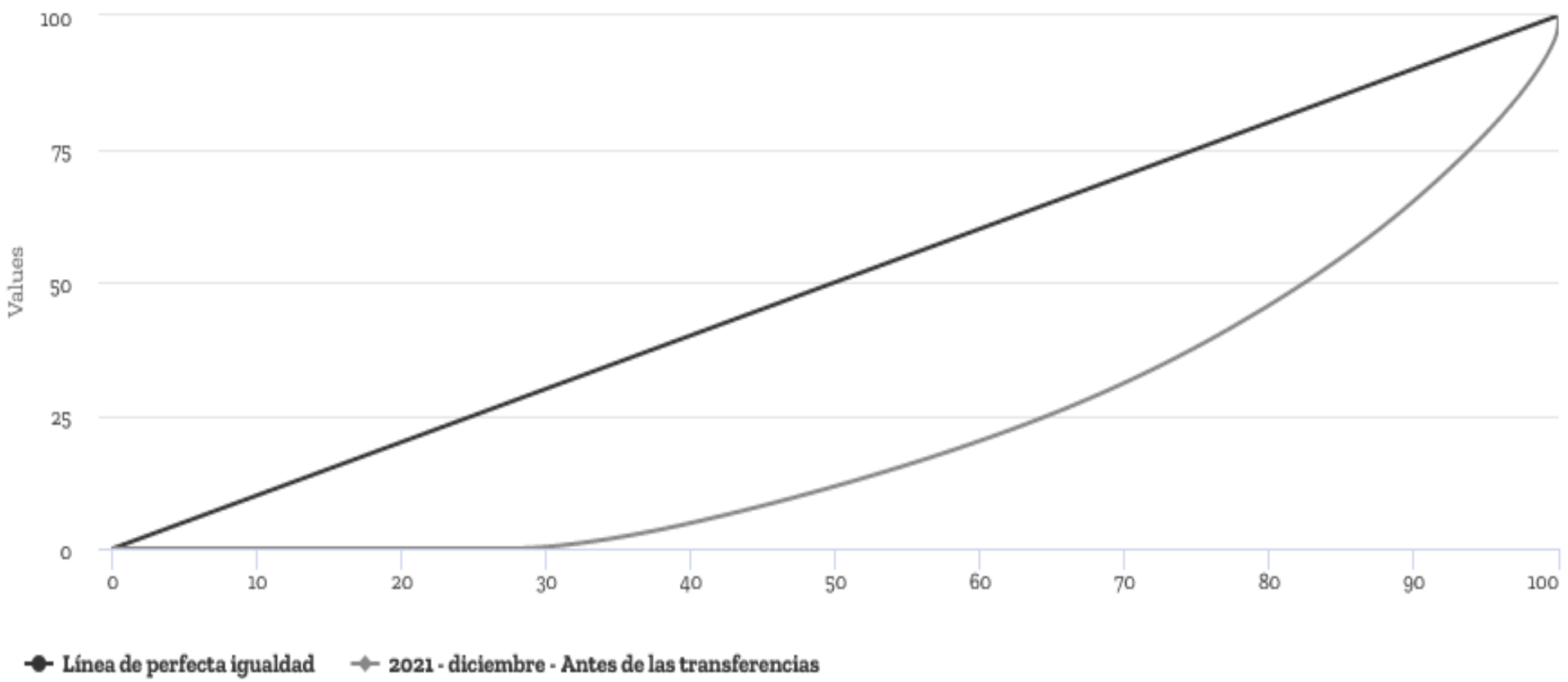
Medidas de Concentración

Curva de Lorenz

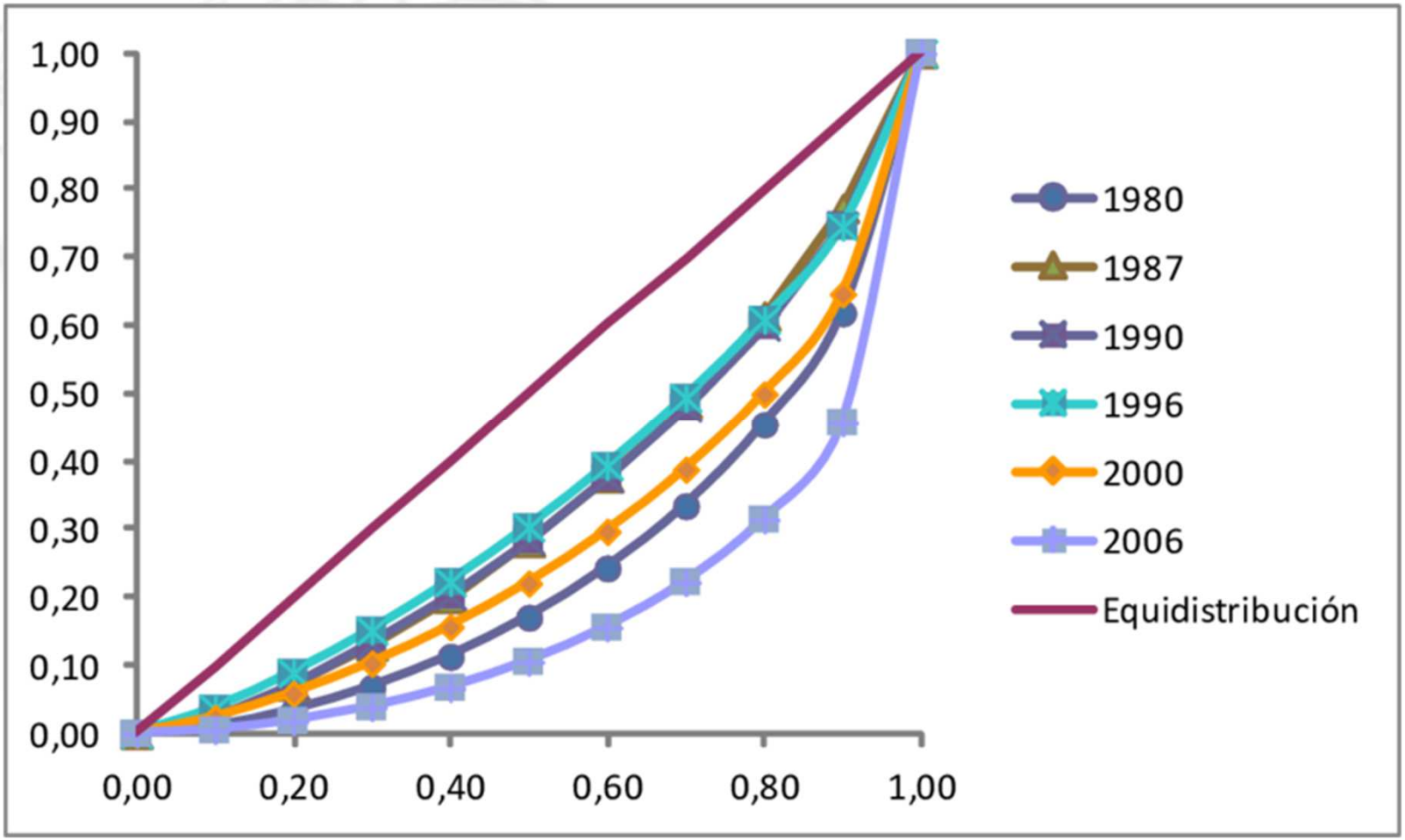
Representaremos gráficamente los pares de puntos (p_i, q_i).



Curva de Lorenz - España



Curva de Lorenz - España



Medidas de Concentración

Curva de Lorenz

La diagonal nos resultará útil para determinar el nivel de concentración de la distribución, pudiendo darse dos casos extremos:

- Concentración mínima: la curva coincide con la diagonal, (se verifica que $p_i=q_i$ para todo i. Estaremos en una situación de máxima equidad.
- Concentración máxima: la curva coincide con los lados del cuadrado, verificándose que $q_i=0$ para $i=1, 2, \ldots, k-1$ y $q_k = 100$. En este caso, no existe equidad alguna en el reparto.

Medidas de Concentración

Índice de Gini

El índice de Gini cuantifica el grado de aproximación existente entre la curva de Lorenz y la línea de equidad. Matemáticamente se expresa como:

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

Medidas de Concentración

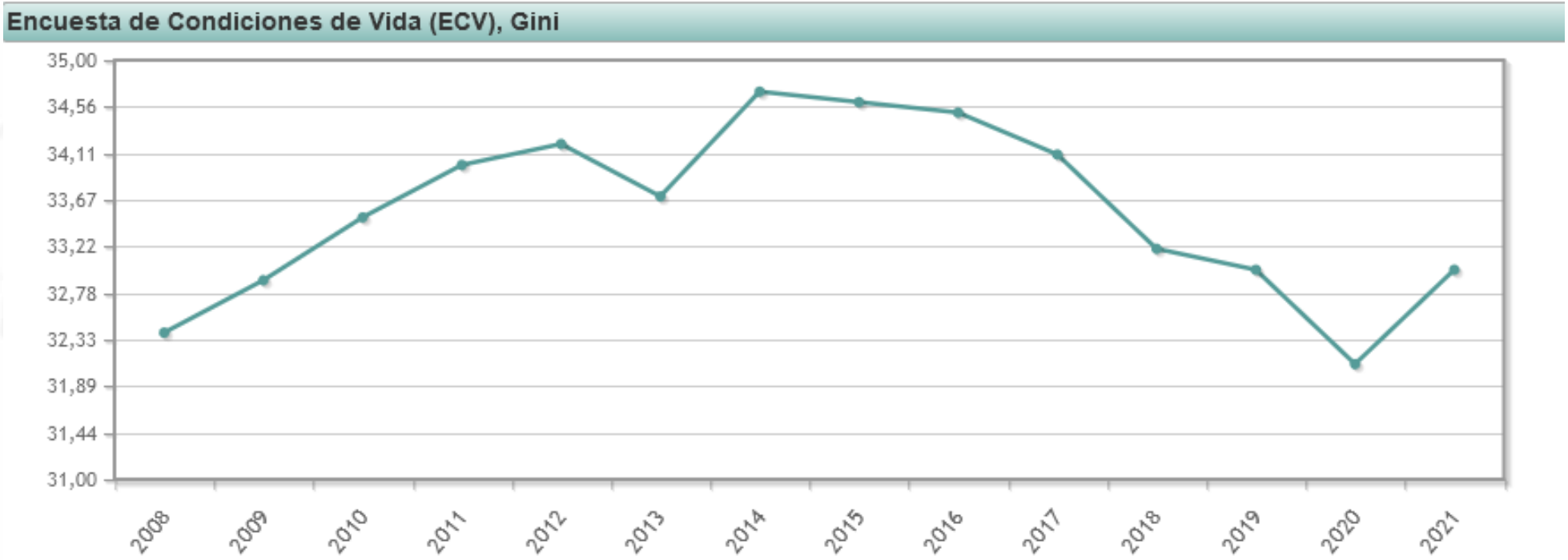
Índice de Gini

El índice de Gini oscila entre 0 y 1. Cuanto más próximo esté su valor a cero, menor será la concentración, es decir, mayor equidad habrá en el reparto de la variable entre los individuos; por el contrario, cuanto más próximo esté a la unidad, mayor será la concentración. Los casos extremos son:

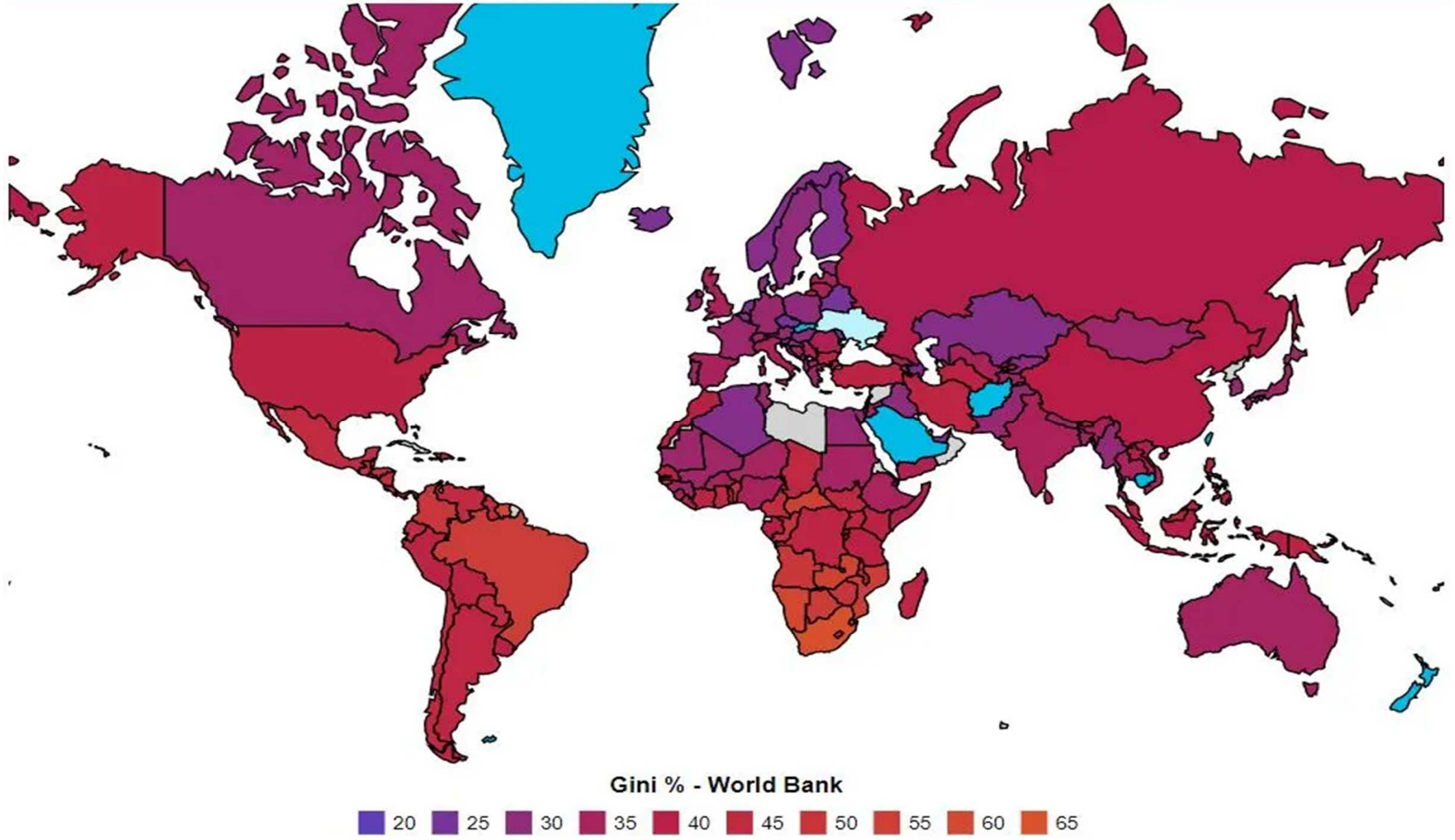
- Concentración mínima (I_G=0), verificándose que p_i=q_i (máxima equidad).

- Concentración máxima (I_G=1), al verificarse que q_i=0 para i=1, 2, …, k-1 y q_k=100 (mínima equidad, máxima desigualdad).

Coef. de Gini España (2008-2021)



Coef. de Gini por Países (2022)



Medidas de Concentración

VEAMOS UN EJEMPLO

Introducción a la Estadística
Grado en Turismo

DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES

Distribuciones de Frecuencias Bidimensionales

Distribuciones Bidimensionales

En este último capítulo del temario entramos en el mundo bidimensional, donde analizaremos relaciones entre dos variables.

En particular estudiaremos en qué grado una variable afecta a otra y aprenderemos a determinar cuál es la causa y cuál es el efecto.

Distribuciones de Frecuencias Bidimensionales

Tabulación de Distribuciones Bidimensionales

- Variable Cuantitativas → Tabla de Correlación
- Variables Cualitativas → Tabla de Contingencia

$X \setminus Y$	y_1	y_2	...	y_s	$n_{i\bullet}$
x_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\bullet}$
...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{s\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet s}$	$N = n_{\bullet\bullet}$

Distribuciones de Frecuencias Bidimensionales

Algunos Conceptos en Distribuciones Bidimensionales

- Frecuencia Absoluta Conjunta (n_{ij}) → $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = N$

- Frecuencia Relativa Conjunta (f_{ij}): $f_{ij} = \frac{n_{ij}}{N}$. Se cumple que:

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}}{N} = \frac{N}{N} = 1$$

Distribuciones de Frecuencias Bidimensionales

Algunos Conceptos en Distribuciones Bidimensionales

- Frecuencias Absolutas Marginales *($n_{i\bullet}$, $n_{\bullet j}$)*. Se verifica que:

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{is} = \sum_{j=1}^{s} n_{ij} \qquad n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{rj} = \sum_{i=1}^{r} n_{ij}$$

$$\sum_{i=1}^{r} n_{i\bullet} = n_{1\bullet} + n_{2\bullet} + \ldots + n_{r\bullet} = N \qquad \sum_{j=1}^{s} n_{\bullet j} = n_{\bullet 1} + n_{\bullet 2} + \ldots + n_{\bullet s} = N$$

Distribuciones de Frecuencias Bidimensionales

Algunos Conceptos en Distribuciones Bidimensionales

- Frecuencias Relativas Marginales ($f_{i\bullet}$, $f_{\bullet j}$). Se verifica que:

$$f_{i\bullet} = \frac{n_{i\bullet}}{N} \qquad f_{\bullet j} = \frac{n_{\bullet j}}{N}$$

Distribuciones de Frecuencias Bidimensionales

Algunos Conceptos en Distribuciones Bidimensionales

Si no hay muchos valores, la tabla de correlación se puede simplificar de la siguiente manera:

x_i	y_j	n_i
x_1	y_1	n_1
x_2	y_2	n_2
...	...	...
x_r	y_s	n_r
		N

Distribuciones de Frecuencias Bidimensionales

Algunos Conceptos en Distribuciones Bidimensionales

Y si la frecuencia es unitaria, entonces podemos prescindir de la columna de frecuencias:

x_i	y_i
x_1	y_1
x_2	y_2
...	...
x_r	y_s

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Distribuciones Marginales

Para analizar el comportamiento individual de cada una de las variables sin considerar lo que hace la otra variable usaremos lo que se conoce como distribuciones marginales:

X		Y	
x_i	$n_{i\bullet}$	y_j	$n_{\bullet j}$
x_1	$n_{1\bullet}$	y_1	$n_{\bullet 1}$
x_2	$n_{2\bullet}$	y_2	$n_{\bullet 2}$
...	...	...	...
x_r	$n_{r\bullet}$	y_s	$n_{\bullet s}$
	N		N

Distribuciones de Frecuencias Bidimensionales

Distribuciones Marginales

A partir de las distribuciones marginales de cada variable podemos calcular medidas de posición y de dispersión:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_{i\bullet}}{N} \qquad S_X^2 = \frac{\sum_{i=1}^{r} \left(x_i - \bar{x}\right)^2 n_{i\bullet}}{N} = \frac{\sum_{i=1}^{r} x_i^2 n_{i\bullet}}{N} - \bar{x}^2$$

$$\bar{y} = \frac{\sum_{j=1}^{s} y_j n_{\bullet j}}{N} \qquad S_Y^2 = \frac{\sum_{j=1}^{s} \left(y_j - \bar{y}\right)^2 n_{\bullet j}}{N} = \frac{\sum_{j=1}^{s} y_j^2 n_{\bullet j}}{N} - \bar{y}^2$$

Distribuciones de Frecuencias Bidimensionales

Distribuciones Condicionadas

Nos permiten analizar el comportamiento de una de las variables sujeto a un valor predefinido de la otra variable.

X	
$x_i / Y = y_j$	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
...	...
x_r	n_{rj}
	$n_{\bullet j}$

Y	
$y_j / X = x_i$	$n_{j/i}$
Y_1	n_{i1}
Y_2	n_{i2}
...	...
Y_s	n_{is}
	$n_{i\bullet}$

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Dependencia Estadística

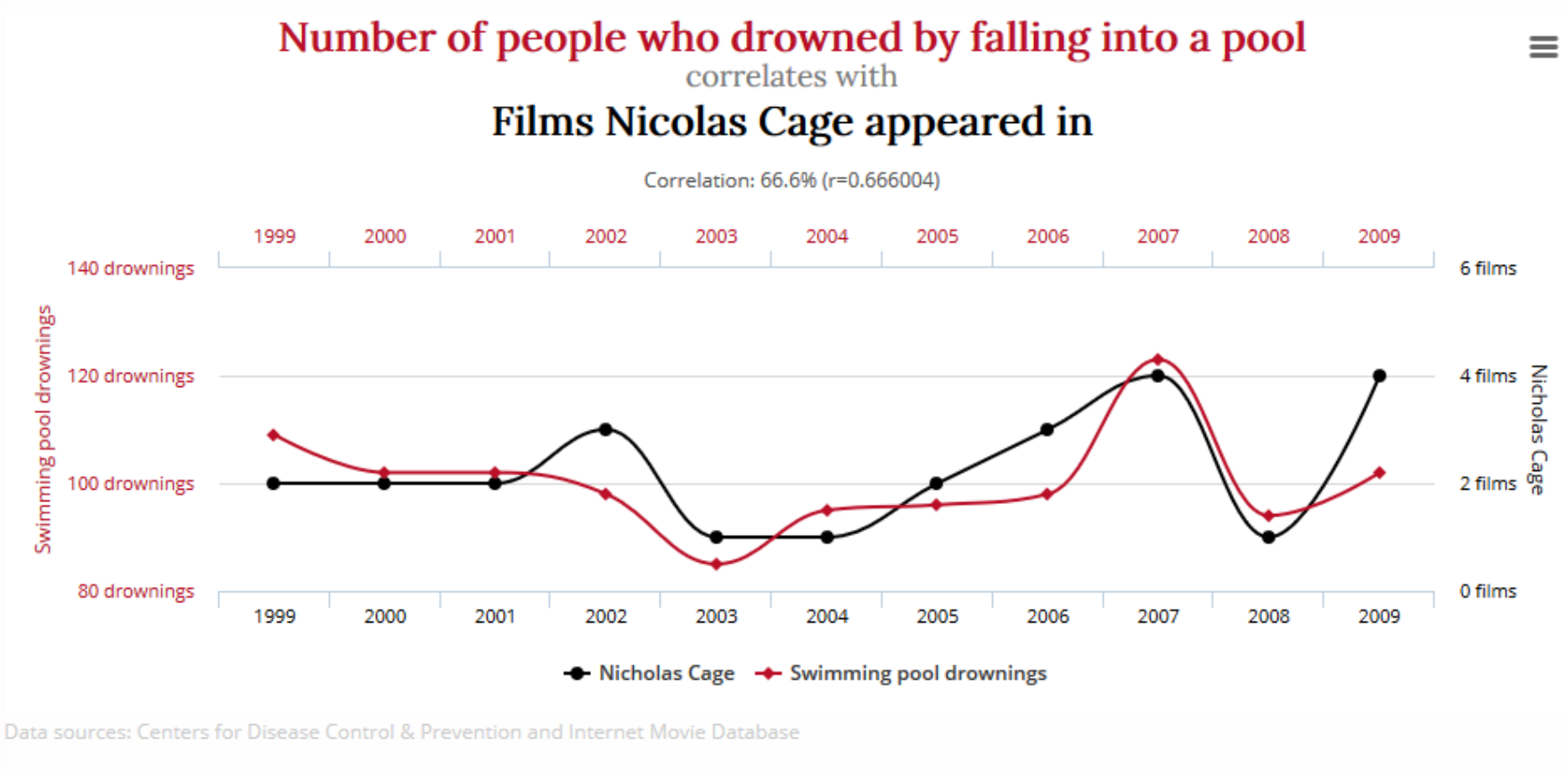
Como decíamos en la clase anterior, nuestro interés en el mundo bidimensional se va a centrar en analizar relaciones entre dos variables. Dicho análisis podemos plantearlo desde dos puntos de vista diferentes:

- Estudiar la intensidad y signo de la relación → Correlación.

- Explicar el comportamiento de una variable a partir del comportamiento de la otra mediante un modelo matemático → Regresión.

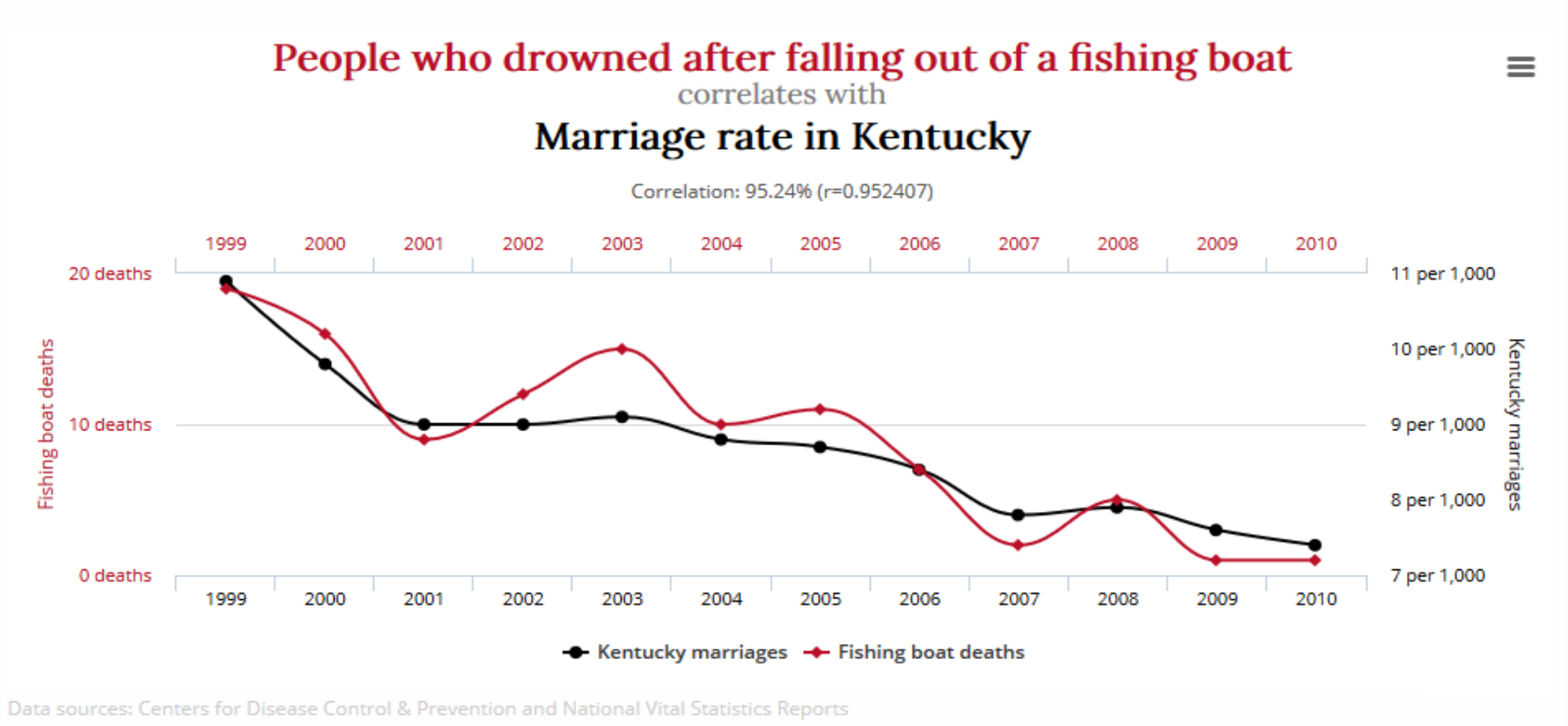
Distribuciones de Frecuencias Bidimensionales

Dependencia Estadística



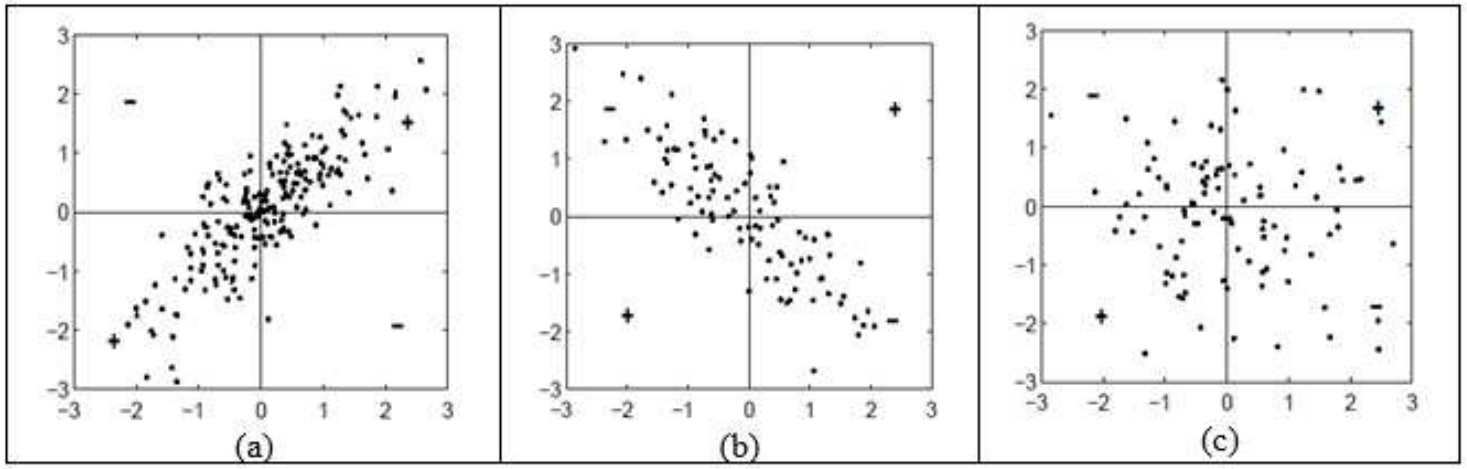
Distribuciones de Frecuencias Bidimensionales

Dependencia Estadística



Distribuciones de Frecuencias Bidimensionales

Dependencia Estadística



Directa **Inversa** **Sin relación**

Distribuciones de Frecuencias Bidimensionales

Causalidad vs. Casualidad

¡Correlación no siempre implica causalidad!

Para que X cause a Y debe verificarse que:

- X precede a Y
- Y no ocurre cuando X no ocurre
- Y ocurre cuando X ocurre

Reflexión: ¿las bodas causan el buen tiempo en primavera?

Distribuciones de Frecuencias Bidimensionales

Covarianza

$$S_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right)n_{ij}}{N}$$

O alternativamente:

$$S_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{ij}}{N} - \bar{x}\bar{y}$$

Distribuciones de Frecuencias Bidimensionales

Covarianza - Propiedades

- Si $S_{XY} > 0$ → Relación directa. Si $S_{XY} < 0$ → Relación inversa. Si $S_{XY} = 0$ → No hay relación.

- **Importante:** si dos variables son independientes, su covarianza es cero, pero el recíproco de esta afirmación **no** siempre es cierto, por lo que si la covarianza entre dos variables es nula, las variables no son necesariamente independientes.

- Siempre se verifica que $S_{XY} = S_{YX}$

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Covarianza - Propiedades

La covarianza es invariable ante cambios de origen pero es sensible ante cambios de escala tal que:

$$U = a+bX$$

$$V = c+dY$$

$$\rightarrow S_{UV} = b \cdot d \cdot S_{XY}$$

Al ser la covarianza sensible a cambios en las unidades de medida necesitamos normalizar de alguna manera su valor. Para ello, utilizaremos el coeficiente de correlación lineal de Pearson.

Distribuciones de Frecuencias Bidimensionales

Coeficiente de Correlación Lineal de Pearson

Se calcula mediante:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

- Oscila siempre entre -1 y +1.
- Si $r_{XY} > 0$ → Relación directa (aceptable a partir de +0,75)
- Si $r_{XY} < 0$ → Relación inversa (aceptable a partir de -0,75)
- Si $r_{XY} = 0$ → No hay relación lineal

Distribuciones de Frecuencias Bidimensionales

Coeficiente de Correlación Lineal - Propiedades

- Si multiplicamos los valores de las variables por una constante, su valor no varía (invariante ante cambios de escala).

- Si dos variables son independientes, su coeficiente de correlación es igual a cero, pero si el coeficiente de correlación es nulo, no significa que las variables sean necesariamente independientes.

Distribuciones de Frecuencias Bidimensionales

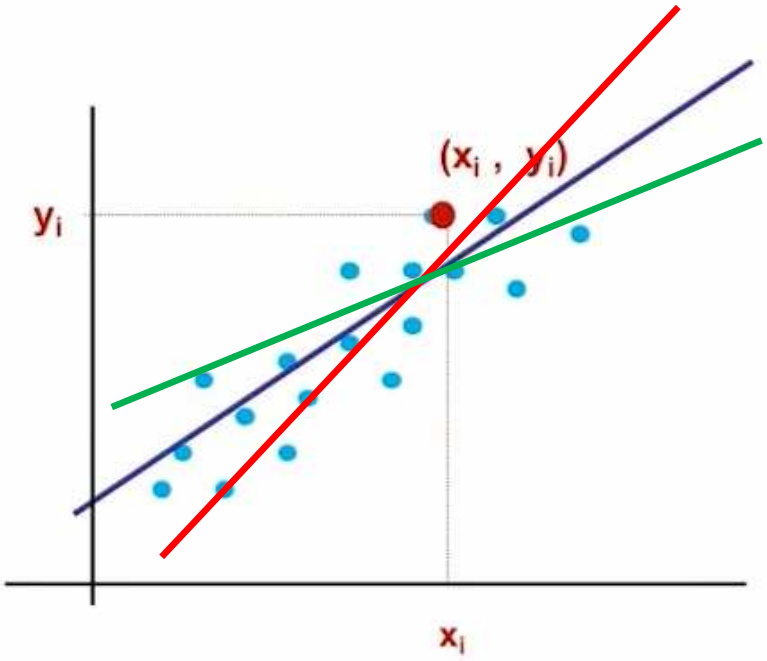
VEAMOS UN EJEMPLO

Introducción a la Estadística
Grado en Turismo

DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión



Pregunta: ¿qué condición debe cumplir la recta que mejor se ajusta a la nube de puntos?

$$Y = a + bX$$

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión

La mejor recta será aquella que minimice SC_e:

$$SC_e = \sum_{i=1}^{k} e_i^2 = \sum_{i=1}^{k} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{k} \left(y_i - (a + bx_i) \right)^2$$

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión

Regresión de Y sobre X → $Y = a + bX$:

$$b = \frac{S_{XY}}{S_X^2}$$

$$a = \bar{y} - b\bar{x} = \bar{y} - \frac{S_{XY}}{S_X^2} \cdot \bar{x}$$

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión

Regresión de X sobre Y → $X = a + bY$:

$$b = \frac{S_{XY}}{S_Y^2}$$

$$a = \bar{x} - b\bar{y} = \bar{x} - \frac{S_{XY}}{S_Y^2} \cdot \bar{y}$$

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión

Dos cuestiones importantes:

1. ¿Qué significado tienen *a* y *b*?

→ *a* es el término constante y representa el valor fijo de Y cuando X es igual a 0.

→ *b* es la pendiente de la recta y representa cuánto aumenta Y por cada unidad que aumenta X. Su signo viene determinado por la covarianza (>0 → Relación directa / <0 → Relación inversa).

Distribuciones de Frecuencias Bidimensionales

Recta de Regresión

Dos cuestiones importantes:

2. ¿Para obtener la recta de X sobre Y, puedo despejar la recta de Y sobre de X?

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Bondad del Ajuste - Coeficiente de Determinación R^2

El Coeficiente de Determinación mide la proporción de variabilidad de la variable dependiente respecto a su media que es explicada por el modelo de regresión.

$$R^2 = r_{XY}^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}$$

Alternativamente también se puede obtener como cociente entre las varianzas de Y e Y*.

R^2 oscila entre 0 y +1, siendo considerado un valor aceptable (la recta es representativa de la relación entre X e Y) a partir de 0,75.

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Predicción

Podemos realizar previsiones para la variable dependiente utilizando un valor dado para la variable independiente. En particular, podemos:

- Pronosticar los valores de la variable dependiente a partir de valores de la variable independiente que pertenecen al intervalo de variación de los datos observados → **Interpolación**.
- Predecir valores de la variable dependiente a partir de valores de la variable independiente que estén situados fuera de dicho intervalo → **Extrapolación.**

Distribuciones de Frecuencias Bidimensionales

Predicción

A la hora de realizar predicciones, debemos tener en cuenta que su fiabilidad dependerá de dos factores:

- La calidad del ajuste: nuestras predicciones serán más fiables cuanto mejor sea el ajuste, es decir, cuanto mayor sea el valor del coeficiente R^2.

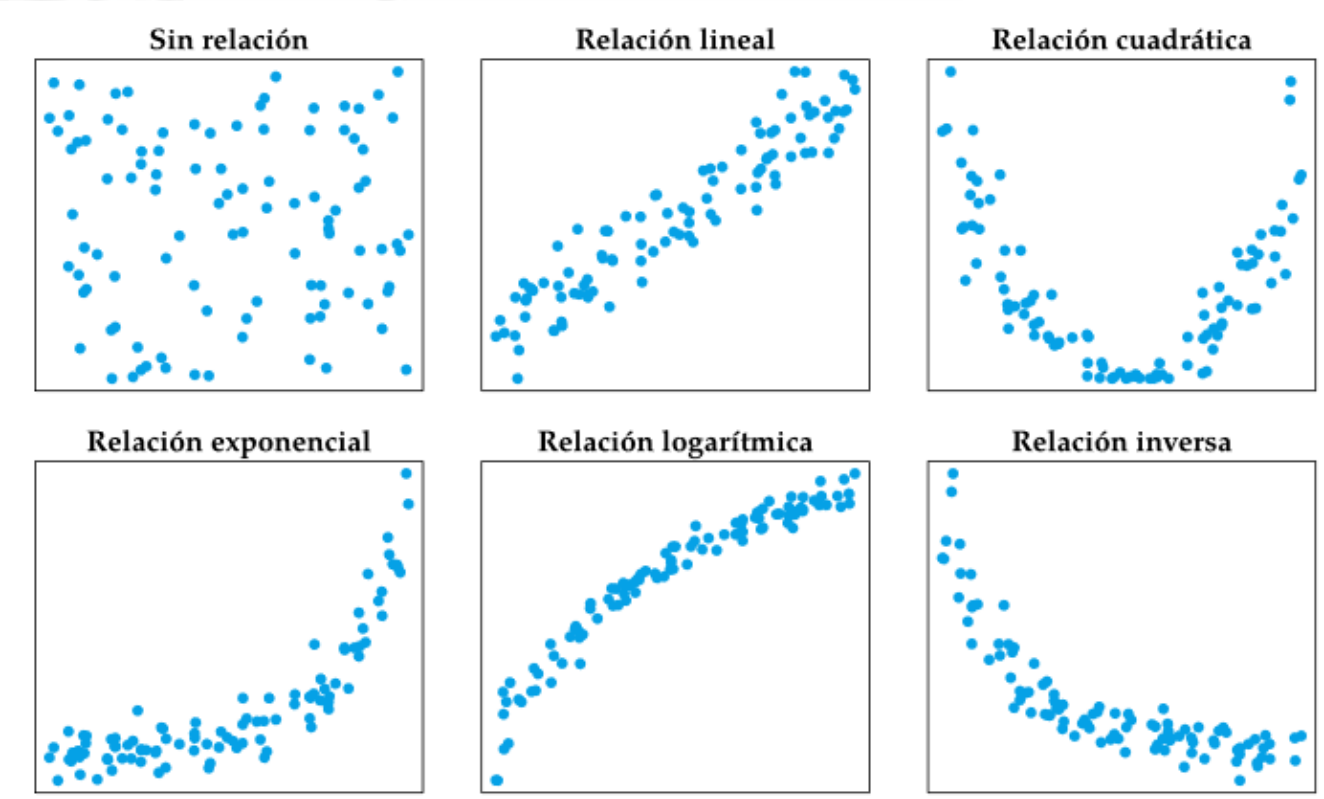
- Los valores de la variable independiente: la fiabilidad de la predicción disminuirá a medida que nos alejemos del rango que comprende a los datos de partida, ya que desconocemos cómo es la relación entre las variables a partir de determinados valores.

Distribuciones de Frecuencias Bidimensionales

VEAMOS UN EJEMPLO

Distribuciones de Frecuencias Bidimensionales

Relaciones No Lineales



¡¡¡HEMOS TERMINADO EL TEMARIO!!!