

## INTRODUCCIÓN A LA ESTADÍSTICA.

**CURSO 2.000-2.001. FEBRERO. Código de carrera 43. Código de asignatura 203.**

### Preguntas teórico-prácticas

**1.-** Dada la siguiente distribución unidimensional

$x_i$	$n_i$
-3	1
-2	5
-1	1
+1	1
+2	5
+3	1

¿Qué medida de posición se debe utilizar? Razone la respuesta.

**Respuesta.-**

Al tratarse de una distribución simétrica, la media y la mediana coinciden (ambas valen cero). Sin embargo es bimodal, siendo -2 y 2 las modas. Por tanto, la medida de posición que se debe utilizar es el **la moda** puesto que es la que más información proporciona acerca de la distribución de la población.

**2.-** Índice de concentración de Gini. Interpretación.

**Respuesta.-**

Consideremos una población de  $N$  individuos y la variable estadística  $X = \{x_i, n_i\}$ ,  $i = 1, 2, 3, \dots, r$ , donde  $x_i$  es la renta correspondiente a cada uno de los  $n_i$  individuos, siendo  $N = \sum_{i=1}^r n_i$ . Supongamos además que  $x_1 < x_2 < \dots < x_r$ . Llamemos  $u_i = \sum_{j=1}^i x_j \cdot n_j$  es decir, la renta

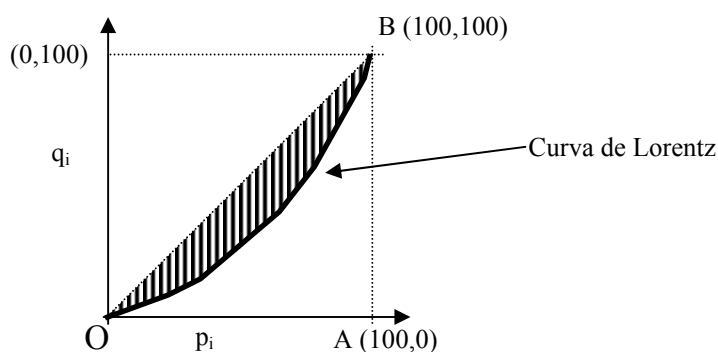
total poseída por los individuos cuya renta es menor o igual que  $x_i$ . Sea  $p_i = \frac{N_i^\uparrow}{N} \cdot 100$ , es decir, el porcentaje de individuos poseedores de la renta  $u_i$  ( $N_i^\uparrow$  es la frecuencia acumulada ascendente, es decir, el número de individuos poseedores de la renta  $u_i$ ); sea  $q_i = \frac{u_i}{u_r} \cdot 100$  es decir, el porcentaje de renta poseída por los  $N_i^\uparrow$  individuos anteriores. El índice de Gini es:

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

En la figura se

observa que  $\sum_{i=1}^{r-1} (p_i - q_i)$  es aproximadamente igual al área comprendida entre la diagonal  $OB$  y la curva de Lorentz, mientras que  $\sum_{i=1}^{r-1} p_i$  es

aproximadamente igual al área del triángulo  $OAB$ . El índice de Gini es por lo tanto aproximadamente igual a la razón entre ambas áreas.



(Nota: Obsérvese que al ser  $\frac{u_i}{N_i^{\uparrow}} \leq \frac{u_r}{N} \Rightarrow \frac{u_i}{u_r} \leq \frac{N_i^{\uparrow}}{N} \Leftrightarrow q_i \leq p_i$ )

**3.-** Relación existente entre la varianza de la variable dependiente, la varianza explicada por la regresión y la varianza residual. Significado de cada una de ellas.

**Respuesta.-**

Supongamos que la recta de regresión de Y/X es  $y = a + bx$ . La relación que se pide es:

$S_y^2 = S_{t_y}^2 + S_{r_y}^2$ , donde  $S_y^2 = m_{02}$  es la varianza de la variable dependiente,  $S_{t_y}^2 = \frac{m_{11}^2}{m_{20}}$  es la

varianza explicada por la regresión (varianza de la variable  $a + bx_i$ ) y  $S_{r_y}^2 = m_{02} - \frac{m_{11}^2}{m_{20}}$  la

varianza residual (varianza de la variable  $r_i = y_i - a - bx_i$ ).

**4.-** Cite y formule cuatro índices complejos de precios ponderados. ¿Qué propiedades cumplen y cuáles no cumplen estos índices?

**Respuesta.-**

Índice de Laspeyres  $P_L = \frac{\sum_i p_{it} q_{i0}}{\sum_i p_{i0} q_{i0}}$ ; índice de Paasche:  $P_P = \frac{\sum_i p_{it} q_{it}}{\sum_i p_{i0} q_{it}}$ ; índice de

Edgeworth:  $P_E = \frac{\sum_i p_{it} (q_{i0} + q_{it})}{\sum_i p_{i0} (q_{i0} + q_{it})}$ ; índice de Fisher:  $P_F = \sqrt{P_L \cdot P_P}$

(generalmente suelen presentarse multiplicados por 100)

Las propiedades se resumen en el siguiente cuadro:

	Laspeyres	Paasche	Edgeworth	Fisher
Existencia	Si	Si	Si	Si
Identidad	Si	Si	Si	Si
Inversión	No	No	Si	Si
Circular	No	No	Si	Si
Proporcionalidad	Si	Si	Si	Si

## Problemas

1.- Una cadena hotelera tiene cinco hoteles de diferente número de plazas cada uno de ellos. Los ingresos totales (referidos a un cierto período) y el rendimiento por habitación de cada hotel son los siguientes:

Hoteles	Ingresos totales hotel (en pesetas)	Rendimiento medio (en pesetas)
1	2.000.000	10.000
2	3.600.000	9.000
3	2.500.000	5.000
4	2.400.000	8.000
5	1.800.000	1.200

Halle el rendimiento medio por habitación para el total de los establecimientos de la cadena hotelera.

### Solución.-

Llamamos  $x_i$  al rendimiento medio de la habitación del hotel  $i$  y  $n_i$  al número de habitaciones que tiene el hotel  $i$ , entonces el ingreso total del hotel  $i$  sería  $x_i \cdot n_i$ .

Construimos la tabla:

Hoteles	Ingresos totales hotel (en pesetas) = $x_i \cdot n_i$	Rendimiento medio (en pesetas) = $x_i$	$n_i = \frac{x_i \cdot n_i}{x_i}$
1	2.000.000	10.000	200
2	3.600.000	9.000	400
3	2.500.000	5.000	500
4	2.400.000	8.000	300
5	1.800.000	1.200	1500
<b>Totales:</b>	<b>12.300.000</b>		<b>2900</b>

Así pues, el rendimiento medio por habitación para el total de los establecimientos de la

cadena hotelera sería 
$$\frac{\sum_{i=1}^5 x_i \cdot n_i}{\sum_{i=1}^5 n_i} = \frac{12300000}{2900} \cong 4241,38 \text{ pts.}$$

2.- El propietario de un piso pactó con su inquilino un alquiler, el 1-1-1.995, de 120.000 pts. mensuales, con cláusula de revisión anual en base a los incrementos experimentados por el índice de precios de consumo anual. Se adjuntan los índices, facilitados por el organismo competente a 31 de diciembre y referidos al año base 1.992. ¿Cuál será el alquiler a pagar los años 1.996, 1.997 y 1.998?

Años	Índice de precios de consumo
1.994	111,9
1.995	116,7
1.996	120,5
1.997	122,9

### Solución.-

Cambiamos en primer lugar los índices de precios de la base 1992 a la base 1994:

Años	Índice de precios base 1992	Índice de precios base 1994
1.994	111,9	100,0
1.995	116,7	104,3
1.996	120,5	107,7
1.997	122,9	109,8

Cuando finalice 1995, se publicará el índice de precios (104,3) y en ese momento el alquiler para 1996 deberá ser una cantidad  $x_{96}$  que, deflactada dé 120000 pts:

$$\frac{x_{96}}{104,3} \cdot 100 = 120000 \Rightarrow x_{96} = \frac{120000 \cdot 104,3}{100} = 125147. \text{ Análogamente se obtienen:}$$

$$x_{97} = \frac{120000 \cdot 107,7}{100} = 129223 \text{ y } x_{98} = \frac{120000 \cdot 109,8}{100} = 131796$$



## INTRODUCCIÓN A LA Estadística (ECONOMÍA). - EXAMEN PRINCIPAL. CURSO 2.000-2.001. SEPTIEMBRE.

### Preguntas teóricas

1.- En un determinado país en período preelectoral, en el que existen cuatro partidos que presentan candidatura a la Presidencia, se realiza una encuesta sobre la tendencia de voto. Se obtienen los siguientes resultados: el 30% de la población votará A, el 25% B, el 5% C y el 20% D, existiendo un 20% de votantes indecisos. ¿Qué medida de posición es la más representativa para esta distribución? Razone la respuesta.

#### **Respuesta.-**

La tendencia de voto es una característica poblacional **cualitativa** y en este caso la única medida de posición es la **moda**.

2.- Coeficiente de variación de Pearson. ¿Es invariante por cambio de escala? (¿y de origen? Razone la respuesta.

#### **Respuesta.-**

Sea  $x_i$  una variable de frecuencia  $n_i$ ,  $i = 1, 2, \dots, r$ , y  $\sum_{i=1}^r n_i = N$ . Efectuar un

cambio de escala consiste en multiplicar  $x_i$  por una constante (positiva) y efectuar un cambio de origen consiste en sumar a  $x_i$  una constante. Sea pues  $x'_i = px_i + q$

$$\text{La media de } x'_i: \overline{x'_i} = \frac{1}{N} \sum_{i=1}^r (px_i + q)n_i = p \cdot \frac{1}{N} \sum_{i=1}^r x_i n_i + q \cdot \frac{1}{N} \sum_{i=1}^r n_i = p\overline{x} + q$$

(es decir, **la media se ve afectada por el mismo cambio de origen y de escala efectuado en la variable**)

$$\text{La varianza de } x'_i: S_{x'}^2 = \frac{1}{N} \sum_{i=1}^r (x'_i - \overline{x'_i})^2 = \frac{1}{N} \sum_{i=1}^r (px_i - p\overline{x})^2 = p^2 \frac{1}{N} \sum_{i=1}^r (x_i - \overline{x})^2 = p^2 \cdot S_x^2$$

es decir, es invariante ante el cambio de origen pero no ante el cambio de escala. La desviación típica será:  $S_{x'} = p \cdot S_x$ .

$$\text{Por tanto, el coeficiente de variación: } CV_{x'} = \frac{S_{x'}}{\overline{x'}} = \frac{p \cdot S_x}{p\overline{x} + q}.$$

Tenemos los siguientes casos:

-si sólo efectuamos un cambio de origen ( $p = 1$ ,  $q \neq 0$ ), entonces  $CV_{x'} = \frac{S_x}{\overline{x} + q} \neq CV_x$ , luego **no es invariante por un cambio de origen**.

- si sólo efectuamos un cambio de escala ( $q = 0$ ) entonces  $CV_{x'} = \frac{S_{x'}}{\overline{x'}} = \frac{p \cdot S_x}{p\overline{x}} = \frac{S_x}{\overline{x}} = CV_x$ , luego **es invariante por cambio de escala**.



**3.-** Campo de variación del coeficiente de correlación lineal. Interpretación del valor nulo de este coeficiente, ¿qué se puede decir de las rectas de regresión en este caso? Razone la respuesta.

**Respuesta.-**

El coeficiente de correlación lineal  $R = \frac{m_{11}}{\sqrt{m_{20} \cdot m_{02}}}$  cumple que  $-1 \leq R \leq 1$ .

Las rectas de regresión:

$$\text{- de Y/X: } y - a_{01} = \frac{m_{11}}{m_{20}} (x - a_{10})$$

$$\text{- de X/Y: } x - a_{10} = \frac{m_{11}}{m_{02}} (y - a_{01})$$

Si  $R = 0$ , no existe correlación entre las variables y, por ser  $m_{11} = 0$ , las rectas de regresión serían:

$$\text{- de Y/X: } y - a_{01} = 0$$

$$\text{- de X/Y: } x - a_{10} = 0$$

que son dos rectas perpendiculares, respectivamente paralelas a los ejes de coordenadas y que se cortan en el punto  $(a_{10}, a_{01}) = (\bar{x}, \bar{y})$

**4.-** Definición de variación estacional de una serie temporal. Cite los métodos para la determinación de dicha componente.

**Respuesta.-**

Variación estacional es aquella variación periódica, cuyo periodo es menor o igual a un año.

Si las componentes de la serie temporal tienen carácter multiplicativo se usa el método de la razón a la media móvil para determinar la componente estacional.

Y si las componentes tienen carácter aditivo, se usa el método de la tendencia por ajuste mínimo cuadrático.

## **Problemas**

**1.-** Los hermanos X e Y deciden realizar una donación de distinto importe a sus cuatro sobrinos. El Sr. X dona las siguientes cantidades a cada uno de ellos: 600.000 pts, 600.000 pts, 400.000 pts y 400.000 pts. En cuanto a la realizada por el Sr. Y se desglosa así: 1.200.000 pts, 1.300.000 pts, 1.400.000 pts y 1.100.000 pts. ¿Cuál de los repartos es más equitativo? Justifique la respuesta en base a la teoría estadística estudiada.

**Solución.-**

Puesto que se trata de un problema de distribución de rentas, calcularemos el índice de Gini en cada caso:

$x_i$	$n_i$	$N_i$	$p_i$	$x_i \cdot n_i$	$u_i$	$q_i$
400000	2	2	50	800000	800000	40
600000	2	4	100	1200000	2000000	100

$$\text{en este caso } I_{G_x} = \frac{50 - 40}{50} = 0,2$$

$x_i$	$n_i$	$N_i$	$p_i$	$x_i \cdot n_i$	$u_i$	$q_i$	$p_i - q_i$
1100000	1	1	25	1100000	1100000	22	3
1200000	1	2	50	1200000	2300000	46	4
1300000	1	3	75	1300000	3600000	72	3
1400000	1	4	100	1400000	5000000	100	0

y en este caso:  $I_{G_Y} = \frac{3+4+3}{25+50+75} = \frac{10}{150} \cong 0,67$ .

Por tanto es más equitativo el reparto del Sr. X, al ser  $I_{G_X} < I_{G_Y}$

**2.-** La Universidad Central de un cierto país ha realizado un estudio de la relación existente entre la exposición a un elemento contaminante y el número de personas que han desarrollado una nueva enfermedad. Esta investigación concluye que sí existe dicha relación, con una recta de regresión estimada de  $y = -2 + 1,2x$ , siendo "y" el porcentaje de personas afectadas, "x" los años de exposición a este elemento y el coeficiente de correlación lineal igual a  $-0,8$ .

**a)** Explíquese el significado de los valores  $-2$  y  $1,2$  en la recta de regresión. **b)** ¿Qué porcentaje de enfermos puede esperarse para personas que han estado en contacto con el elemento contaminante durante 30 años? **c)** Si el coeficiente de correlación lineal hubiera sido igual a 1 ¿podríamos decir que el elemento contaminante es la única causa de la enfermedad?

#### **Solución.-**

(Existe una contradicción en el enunciado ya que el coeficiente de correlación ( $-0,8$ ) y la pendiente de la recta de regresión ( $1,2$ ) deben tener el mismo signo. Corregiremos pues el enunciado suponiendo que el coeficiente de correlación lineal es  $0,8$ ).

**a)** El valor  $-2$  de la recta de regresión es la ordenada en el origen. Carece de significado estadístico pues sería el porcentaje de personas afectadas, expuestas "cero" años al elemento contaminante. Esto implica además que, hasta que no pasen  $\frac{5}{3}$  años (1 año y 8 meses) de

exposición, no comenzará a haber enfermos ya que  $-2 + 1,2 \cdot \frac{5}{3} = 0$

El valor  $1,2$  es la pendiente, esto es, el porcentaje que aumenta el número de personas afectadas, por año de exposición.

**b)** Haciendo  $x = 30$  en la recta de regresión:  $y = -2 + 1,2 \cdot 30 = 34$ , es decir, el 34%.

**c)** Si  $R = 1$  existe entre las variables una dependencia funcional exacta; también el coeficiente de determinación  $R^2 = 1$ , lo cual significa que el tiempo de exposición determina al 100% el porcentaje de enfermos. Ahora bien, no podríamos asegurar que fuese la única causa de la enfermedad por que desconocemos otras características de los enfermos (por ejemplo la edad, o alguna insuficiencia –conocida o desconocida- etc...)