Estadística y análisis de datos

Segundo curso

Profesor: Imanol Mozo

Examen resuelto (8 de enero de 2019)

Donostiako Ekonomia eta Enpresa Fakultatea
EHU

Autor: Beñat Zunzunegi

Gizapedia

gizapedia.hirusta.io

Problema 1

Estudiamos el gasto que se realiza en Navidad tomando como muestra un colectivo de 600 familias de Donostia (300 familias viven en Egia y otras 300 familias viven en el Antiguo):

Datos de Egia:

Gasto (€)	Familias
100-500	80
500-1500	120
1500-2700	100

Datos del Antiguo:

Gasto (€)	Familias
200-700	70
700-1800	150
1800-3000	80

(a) Calcula el consumo medio para las familias de Egia y el Antiguo y contrasta si el resultado obtenido es lo suficientemente representativo para los vecinos de cada barrio.

(b) Si el 25 % de los vecinos que más gasta en Aiete se sitúa por encima de 2300€, ¿qué diferencia en euros se obtiene con el 25 % que más gasta en el Antiguo?

(c) Si un vecino de Egia gasta 2000€ y otro vecino del Antiguo gasta 2100€, ¿quién es más gastador relativamente con respecto a su barrio?

(d) ¿Sería cierto afirmar que el gasto más frecuente en Egia se encuentra dentro del colectivo que está en el segundo intervalo de su tabla?

(a)

Datos de Egia:

x	n	nx	nx^2
300	80	24.000	7.200.000
1000	120	120.000	120.000.000
2100	100	210.000	441.000.000
Sumas	300	354.000	568.200.000

Datos del Antiguo:

x	n	nx	nx^2
450	70	31.500	14.175.000
1250	150	187.500	234.375.000
2400	80	192.000	460.800.000
Sumas	300	411.000	709.350.000

Calculo los consumos medios:

$$\overline{x}(Egia) = \frac{354000}{300} = 1180€; \quad \overline{x}(Antiguo) = \frac{411000}{300} = 1370€$$

Para considerar si las medias son representativas para el consumo de los vecinos de cada barrio, calculo el coeficiente de variación para los datos de cada barrio. Para ello, calculo la desviación típica:

$$s_x(Egia) = \sqrt{\frac{568200000}{300} - 1180^2} = 708{,}23€$$

$$s_x(Antiguo) = \sqrt{\frac{709350000}{300} - 1370^2} = 698{,}28€$$

Ahora calculo los coeficientes de variación:

$$CV(Egia) = \frac{708{,}23}{1180} \times 100 = 0{,}60$$

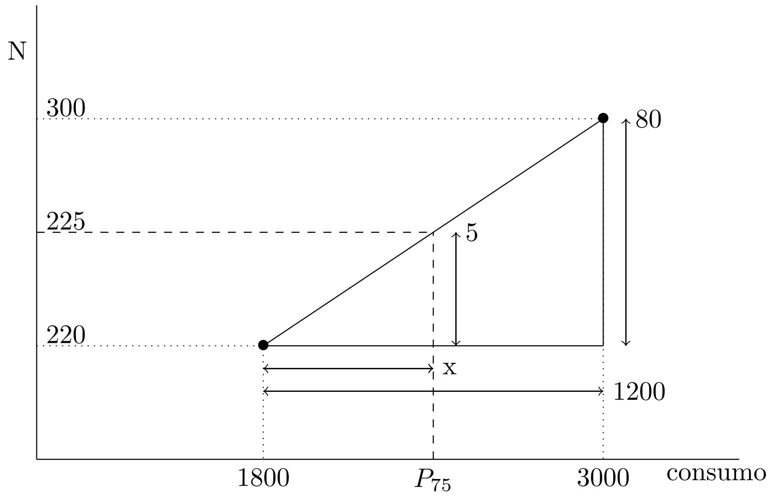
$$CV(Antiguo) = \frac{698{,}28}{1370} \times 100 = 0{,}509$$

Al superar los coeficientes superiores al 50 %, las medias no serían representativas respecto de cada barrio.

(b)

Para poder realizar la comparación hay que calcular el percentil 75, o tercer cuartil, de los datos de consumo del Antiguo.

Como se han recogido datos de 300 vecinos del Antiguo, el percentil 75 será el consumo correspondiente al vecino número $300 \times 0{,}75 = 225$. Dicho vecino se encuentra, tras ir acumulando las frecuencias simples (frecuencias acumuladas: $N: 70, 220, 300$) en el intervalo 1800-3000. Aproximando por interpolación lineal:



$$\frac{x}{5} = \frac{1200}{80} \to x = 75 \to Q_3 = P_{75} = 1800 + 75 = 1875$$

La diferencia de umbral del 25 % de gasto mayor es pues 2300-1875=425€.

(c)

Para comparar relativamente los dos vecinos hay que estandarizar o tipificar los datos:

$$Egia: x = 2000€ \to z = \frac{x - \overline{x}}{s_x} = \frac{2000 - 1180}{708{,}23} = 1{,}15$$

$$Antiguo: x = 2100€ \to z = \frac{x - \overline{x}}{s_x} = \frac{2100 - 1370}{698{,}28} = 1{,}04$$

El vecino de Egia tiene un gastos en relación al conjunto de su barrio superior al del vecino del Antiguo.

(d)

El gasto más frecuente en Egia se corresponde con la moda estadística. Para determinar el intervalo modal en una distribución con intervalos de diferente amplitud, hay que calcular previamente la densidad de cada intervalo:

Intervalo	Frecuencia	Densidad=Frecuencia/Amplitud
100-500	80	80/400=0.2
500-1500	120	120/1000=0.120
1500-2700	100	100/1200=0.083

El intervalo que contiene a la moda es el de mayor densidad, por tanto, la moda está en el intervalo 100-500, por lo que la afirmación del enunciado es falsa.

Problema 2

En la siguiente tabla tenemos los datos de las dietas por desplazamiento que reciben los trabajadores de dos empresas A y B cuando tienen que realizar viajes por razón de su trabajo:

Dietas (€)	**Trabajadores de A**	**Trabajadores de B**
80	**10**	**5**
60	**5**	**15**
20	**5**	**15**
15	**5**	**10**
10	**15**	**5**

Estudie gráfica y analíticamente (es suficiente con un único índice) en cuál de las empresas hay más concentración en el reparto de dietas.

En primer lugar, hay que ordenar la variable estadística de menor a mayor valor. A partir de ahí, se calculan las frecuencias acumuladas y los totales de dietas acumulados, primero en cifras absolutas y luego en porcentaje, para despues construir la curva de Lorenz y calcular un índice de concentración.

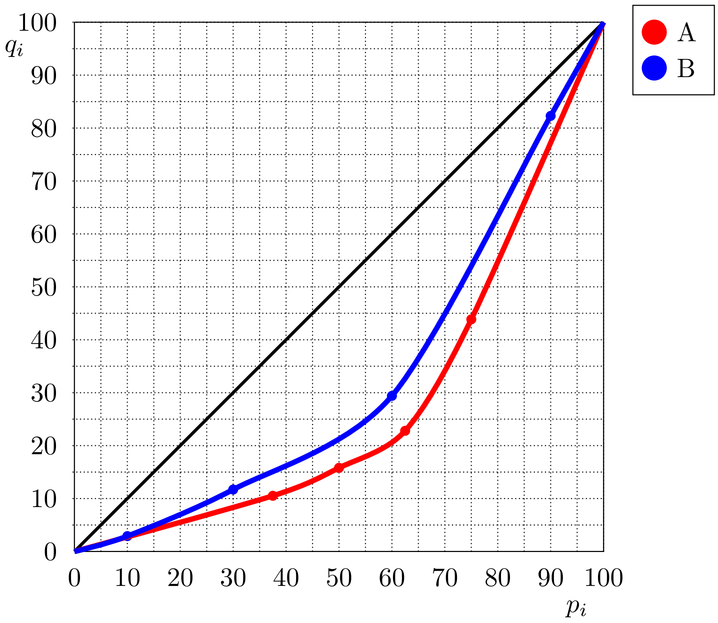
Calculemos los acumulados absolutos, en frecuencias y totales:

x	n_A	n_B	N_A	N_B	$t_A = n_A x$	$t_B = n_B x$	T_A	T_B
10	15	5	15	5	150	50	150	50
15	5	10	20	15	75	150	225	200
20	5	15	25	30	100	300	325	500
60	5	15	30	45	300	900	625	1400
80	10	5	40	50	800	400	1425	1700

Y ahora los acumulados relativos, que conformaran los puntos de la curva de Lorenz:

$p_A = \frac{N_A}{NA}$	$p_B = \frac{N_B}{NB}$	$q_A = \frac{T_A}{TA}$	$q_B = \frac{T_B}{TB}$
37.5	10	10.52	2.9
50	30	15.79	11.7
62.5	60	22.80	29.4
75	90	43.86	82.3
100	100	100	100

Construimos las curvas de Lorenz:



Las dietas de la empresa A se alejan uniformemente más de la diagonal, por lo que muestran una mayor concentración. Confirmamos ahora estos resultados con el índice de Gini:

p_A	p_B	$p_A - q_A$	$p_B - q_B$
37.5	10	26.98	7.1
50	30	34.21	18.3
62.5	60	39.70	30.6
75	90	31.14	7.7
-	-	0	0
225	190	132.03	63.7

$$Gini(A) = \frac{132{,}03}{225} = 0{,}5868$$

$$Gini(B) = \frac{63{,}7}{190} = 0{,}3352$$

La concentración de dietas es mayor en A, tal como habíamos avanzado.

Problema 3

La facturación en miles de euros de una empresa durante los últimos tres años ha sido la siguiente:

Año	**I TRIM.**	**II TRIM.**	**III TRIM.**	**IV TRIM.**
2016	**9**	**9**	**16**	**10**
2017	**9**	**14**	**16**	**13**
2018	**12**	**14**	**20**	**14**

Además, sabemos que los índices de variación estacional (IVE) son los siguientes:

I TRIM.	**II TRIM.**	**III TRIM.**	**IV TRIM.**
0.821	**0.964**	**?**	**0.898**

(a) ¿Cuál es el IVE del tercer trimestre?

(b) Determina la serie desestacionalizada.

(c) Determina la función analítica que permita hacer predicciones.

(d) ¿Cómo es la tendencia a largo plazo? ¿Qué porcentaje de la varianza de la facturación anual está explicada por la tendencia?

(e) Determina la facturación esperada para los cuatro trimestres de 2019.

(a)

Si los IVEs siguen un modelo multiplicativo, su suma debe ser 4. Por tanto,

$$IVE_{III\ trim} = 4 - 0{,}821 - 0{,}964 - 0{,}898 = 1{,}317$$

Indica que en promedio las ventas suben un 31.7 % durante el tercer trimestre, respecto al resto de trimestres.

(b)

Para calcular la serie desestacionalizada, unicamente tenemos que dividir las ventas totales de cada periodo entre el IVE correspondiente. Por ejemplo, para el primer trimestre de 2016 las ventas desestacionalizadas serían: 9/0.821=10.962. Del mismo modo obtenemos el resto de ventas desestacionalizadas:

Año	I TRIM.	II TRIM.	III TRIM.	IV TRIM.
2016	10.962	9.336	12.148	11.135
2017	10.962	14.522	12.148	14.476
2018	14.616	14.522	15.186	15.590

(c)

Asignamos a los trimestres los valores 0, 1, 2, ... (variable X) y la serie desestacionalizada (variable Y) y calculamos la línea de regresión de X sobre Y:

X	Y	XY	X^2
0	10.962	0	0
1	9.336	9.336	1
2	12.148	24.296	4
3	11.135	33.405	9
4	10.962	43.848	16
5	14.522	72.610	25
6	12.148	72.888	36
7	14.476	101.332	49
8	14.616	116.928	64
9	14.522	130.698	81
10	15.186	151.680	100
11	15.590	171.490	121
Suma: 66	Suma: 155.603	Suma: 928.691	Suma: 506

$\hat{T} = a + bX$ (T: tendencia)

$$b = \frac{s_{xy}}{s_x^2} = \frac{\frac{928{,}691}{12} - \frac{66}{12} \cdot \frac{155{,}03}{12}}{\frac{506}{12} - \left(\frac{66}{12}\right)^2} = 0{,}5096$$

$$a = \overline{y} - b\overline{x} = \frac{155{,}03}{12} - 0{,}5096 \cdot \frac{66}{12} = 10{,}1641$$

$\hat{T} = 10{,}1641 + 0{,}5096X$ (T: tendencia)

$\hat{Y} = (10{,}1641 + 0{,}5096X) \times IVE$, $\hat{Y}$: pronóstico de ventas

Problema 4

En la siguiente tabla se muestran los precios por unidad en euros durante el periodo 2000-2005 de las tres materias primas, A, B y C, que se utilizan en una fábrica:

Año	A	B	C
2000	4.43	5.3	6.32
2001	4.23	5.5	7.35
2002	5.56	4.2	7.23
2003	6.34	5.8	8.53
2004	7.22	4.0	6.23
2005	8.11	5.1	15.23

(a) Calcula los índices sinples de precio en el priodo 2000-05 para la materia prima C. ¿En qué proporción ha variado el precio de C entre los años 2000 y 2005?

(b) De nuevo para el precio de C, calcula las tasas de variación interanual y la tasa de variación media en el perido 2000-05.

(c) Sabemos que en el 2000 las cantidades utilizadas fueron las siguientes: 1004 unidades de A, 345 de b y 902 de C. Teniendo en cuenta esto, calcula el índice de Laspeyres de precios para 2005 con periodo de referencia 2000.

(d) Da una aproximación del índice de Laspeyres de precios para 2005 con base en 2003.

(e) Si sabemos que el índice dePaasche de cantidades para 2005 con base en 2000 es 115.67 %, ¿cuál sería el índice de cantidades para ese mismo año con periodo de referencia 2000? Interpreta el resultado.

Problema 5

Tenemos 8 bolas blancas y 4 negras en una urna y extraemos 3 bolas al azar.

(a) Determina el espacio muestral del experimento y calcula la probabilidad de los sucesos elementales.

(b) Define una variable aleatoria bidimensional (x,y) que cuente el número de bolas blancas y negras.

(c) Calcula su función de cuantía.

(d) Estudia la independencia de X e Y.

(a)

Elementos del espacio muestral	Probabilidad
0 blancas	$\frac{4}{12} \times \frac{3}{11} \times \frac{2}{10} = 0{,}018$
1 blanca	$\frac{8}{12} \times \frac{4}{11} \times \frac{3}{10} \times \frac{3!}{2!1!} = 0{,}218$
2 blancas	$\frac{8}{12} \times \frac{7}{11} \times \frac{4}{10} \times \frac{3!}{1!2!} = 0{,}509$
3 blancas	$\frac{8}{12} \times \frac{7}{11} \times \frac{6}{10} = 0{,}254$
	1

(b) y (c)

X: Blancas (↓) / Y: Negras (→)	Y=0	Y=1	Y=2	Y=3	Total
X=0	0	0	0	0.018	0.018
X=1	0	0	0.218	0	0.218
X=2	0	0.509	0	0	0.509
X=3	0.254	0	0	0	0.254
Total	0.254	0.509	0.218	0.018	1

(d)

X e Y son independientes cuando las probabilidades conjuntas o compuestas son el producto de las probabilidades simples o marginales, para todos los valores de X e Y. Veamos para la primera celda:

$$P[X = 0, Y = 0] = P[X = 0] \times P[Y = 0]?$$
$$0 \neq 0{,}018 \times 0{,}254$$

Ya para la primera celda no se cumple la condición por lo que las variables X e Y no son independientes. Al contrario, están perfectamente correlacionadas porque existe una relación funcional entre ellas: $Y = 3 - X$.