# ESTADISTICA Y ANALISIS DE DATOS

## Exámenes resueltos

## Curso 2020-2021

## Facultad de Economía y Empresa. Donostia EHU-UPV

## Profesores: Conchi Oruezabal, Irati Labaien, Jose Mari Sarasola y Lorea Mendiola

Autor: Beñat Zunzunegi

Disponible y periódicamente actualizado, junto con otros exámenes de las misma asignatura en:
https://gizapedia.hirusta.io/examen-estadistica-y-analisis-de-datos-2020-21/

Gizapedia

gizapedia.hirusta.io

ESTADISTICA Y ANALISIS DE DATOS

Profesores: Conchi Oruezabal, Irati Labaien, Jose Maria Sarasola y Lorea Mendiola

Fecha: 14 de enero de 2021, 15:00

Duración: 120 min

## Problema I *(3 puntos)*

Una muestra de 100 contribuyentes recoge la siguiente información sobre impuestos pagados (X) y renta (Y), en cientos de euros:

| $X(\downarrow) - Y(\rightarrow)$ | 6-10 | 10-20 | 20-40 |
|---|---|---|---|
| 0-2 | 20 | 5 | 0 |
| 2-4 | 0 | 30 | 20 |
| 4-6 | 0 | 5 | 20 |

**Tareas a realizar:**

(a) Realizar el histograma de la distribución condicionada de la cantidad de impuestos pagada (X) por la renta de los contribuyentes del intervalo 10-20. Calcular los valores de la media, mediana y moda de esa distribución condicionada, y utilizando el gráfico, explicar los resultados obtenidos.

(b) Analizar el grado de concentración de la distribución de los contribuyentes según su renta, representando la curva de Lorenz y calculando además el coeficiente de Gini.

(c) Analizar el grado de relación lineal que existe entre X e Y en este colectivo. ($\overline{y} = 19{,}6; s_y^2 = 78{,}64$)

---

(a)

Aislo la distribución condicionada x/y:[10:20] :

| $X/Y : [10, 20]$ | n |
|---|---|
| 0-2 | 5 |
| 2-4 | 30 |
| 4-6 | 5 |

Calculo la media. Para ello, aislo la distribución condicionada x/y:[10:20], tomando como valor de referencia la marca de clase de cada intervalo :

| x | n | nx |
|---|---|---|
| 1 | 5 | 5 |
| 3 | 30 | 90 |
| 5 | 5 | 25 |
| | 40 | 120 |

$$\overline{x} = \frac{\sum nx}{\sum n} = \frac{120}{40} = 3 = 300€$$

Calculo la mediana. 40/2=20. Luego la mediana es el dato número 20, y por tanto, acumulando frecuencias, se encuentra en el intervalo segundo. Por interpolación (F, frecuencia acumulada, y f, frecuencia simple, para cada intervalo):

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} a_i = 2 + \frac{20-5}{30} \times 2 = 3 = 300€$$

Calculo la moda. El intervalo modal es de mayor frecuencia simple, por tanto el segundo.

$$Mo = L_i + \frac{f_i - f_{i-1}}{f_i - f_{i-1} + f_i - f_{i+1}} a_i = 2 + \frac{30-5}{(30-5)+(30{,}5)} \times 2 = 3 = 300€$$

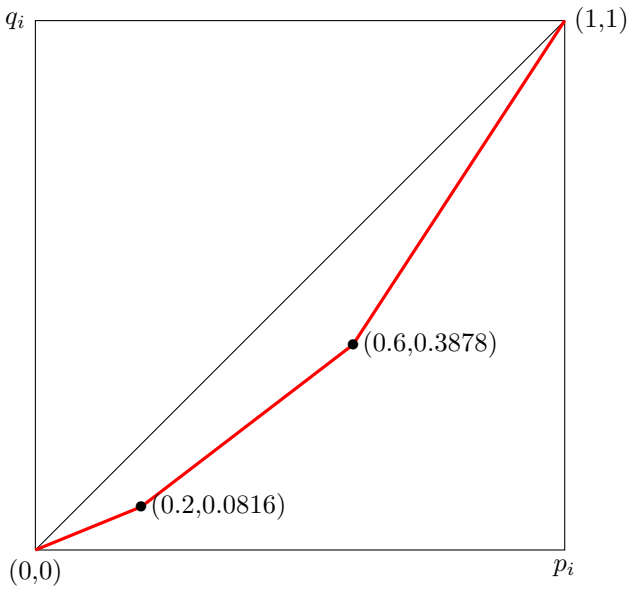
La media aritmètica, mediana y moda coinciden. Por tanto, la distrubicíon es perfectamente simétrica.

(b)

Tomo la distribución marginal de la renta (y), aunque no está muy claro que sea eso lo que pide (hubiese sido más correcto decir distribucón marginal de la renta, en lugar de distribución según renta, que puede dar lugar a equívocos) sumando las frecuencias condicionadas o desagregadas por impuestos pagados (x) (es decir, sumando columnas). Y calculo los totales de renta acumulados estimados para cada intervalo a partir de la marca de clase, para calcular lo valores q. Los valores p los calculo a partir de las frecuencias acumuladas N.

| y | n | ny | ny acum. | N | p=N/100 | q=ny acum./1960 |
|---|---|---|---|---|---|---|
| 8 | 20 | 160 | 160 | 20 | 0.2 | 0.0816 |
| 15 | 40 | 600 | 760 | 60 | 0.6 | 0.3878 |
| 30 | 40 | 1200 | 1960 | 100 | 1 | 1 |
| | 100 | 1960 | | | | |

Construyo la curva de Lorenz, a partir de los puntos (p,q):




Interpretación: por ejemplo, el 20 % de los contribuyentes con menor renta acumula el 8.16 % de la renta total.

Calculo el coeficiente de Gini, a partir de las diferencia p-q y de la suma de los valores p:

| p | p-q |
|---|---|
| 0.2 | 0.2-0.0816=0.1184 |
| 0.6 | 0.6-0.3878=0.2122 |
| 0.8 | 0.3306 |

$$G = \frac{0{,}3306}{0{,}8} = 0{,}41$$

No se puede interpretar el valor del coeficiente sin otras referencias, pero en principio al tomar valores entre 0 y 1, se podría decir que la concentración es intermedia.

(c) Para analizar el grado de relación lineal, se debe calcular el coeficiente de correlación lineal $r_{xy}$ y para ello se deben calcular la covarianza y las desviaciones típicas de las variables. Calculo pues la covarianza y las desviaci(solo de x), ya que la de la variable y viene dada a través de su varianza (ver enunciado), tomando como valores de referencia para las variables $x$ e $y$ las marcas de clase:

| $x$ | $y$ | $n$ | $nx$ | $ny$ | $nxy$ | $nx^2$ |
|---|---|---|---|---|---|---|
| 1 | 8 | 20 | 20 | 160 | 160 | 20 |
| 1 | 15 | 5 | 5 | 75 | 75 | 5 |
| 3 | 15 | 30 | 90 | 450 | 1350 | 270 |
| 3 | 30 | 20 | 60 | 600 | 1800 | 180 |
| 5 | 15 | 5 | 25 | 75 | 375 | 125 |
| 5 | 30 | 20 | 100 | 600 | 3000 | 500 |
| | | 100 | 300 | 1960 | 6760 | 1100 |

$$\overline{x} = \frac{300}{100} = 3 \ ; \ \overline{y} = \frac{1960}{100} = 19{,}6$$

$$s_{xy} = \frac{\sum nxy}{\sum n} - \overline{xy} = \frac{6760}{100} - 3 \times 19{,}6 = 8{,}8$$

$$s_x^2 = \frac{\sum nx^2}{\sum n} - \overline{x}^2 = \frac{1100}{100} - 3^2 = 2 \rightarrow s_x = 1{,}41$$

$$s_y^2 = 78{,}64 \rightarrow s_y = 8{,}86$$

Y por tanto:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{8{,}8}{1{,}41 \times 8{,}86} = 0{,}70$$

Por tanto, la relación lineal entre las dos variables es positiva (al incrementarse la renta suben los impuestos pagados) y esta relación es intensa.

## Problema 2 *(2 puntos)*

Se ha realizado un test a un grupo de alumnos en relación a su desempeño en programación con el lenguaje R, con una puntuación total de 0 a 10. A continuación se recogen las respuestas de una muestra de 8 de esas personas a un item en concreto (c: correcto; i: incorrecto) y las puntuaciones totales obtenidas por dichas personas.

| Individuo | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Item | i | c | i | c | c | i | c | i |
| Total | 2.3 | 3.1 | 4.2 | 5.7 | 6.8 | 7.2 | 8.5 | 9.6 |

Calcular el coeficiente de correlación lineal de Pearson para las dos variables (correlación item-test) e interpretar los resultados en relación a la pertinencia del item.

Fórmulas: $s_x = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$ ; $s_{xy} = \frac{\sum xy}{n} - \overline{xy}$ ; $r_{xy} = \frac{s_{xy}}{s_x s_y}$

---

Utilizo la codificación estándar: incorrecto:0; correcto:1. A partir de esos datos, calculo desviaciones y covarianza.

| $x$ item | $y$ test | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 0 | 2,3 | 0 | 0 | 5,29 |
| 1 | 3,1 | 3,1 | 1 | 9,61 |
| 0 | 4,2 | 0 | 0 | 17,64 |
| 1 | 5,7 | 5,7 | 1 | 32,49 |
| 1 | 6,8 | 6,8 | 1 | 46,24 |
| 0 | 7,2 | 0 | 0 | 51,84 |
| 1 | 8,5 | 8,5 | 1 | 72,25 |
| 0 | 9,6 | 0 | 0 | 92,16 |
| 4 | 47,4 | 24,1 | 4 | 327,52 |

$$\overline{x} = \frac{4}{8} = 0{,}5 \ ; \ \overline{y} = \frac{47{,}4}{8} = 5{,}925$$

$$s_{xy} = \frac{24{,}1}{8} - 0{,}5 \times 5{,}925 = 0{,}05$$

$$s_x = \sqrt{\frac{4}{8} - 0{,}5^2} = 0{,}5$$

$$s_y = \sqrt{\frac{327{,}52}{8} - 5{,}925^2} = 2{,}41$$

$$r_{xy} = \frac{0{,}05}{0{,}5 \times 2{,}41} = 0{,}041$$

La correlación es muy débil, cercana a 0 (y por tanto es irrelevante que sea positiva o negativa). Así pues, el item no está relacionado con el test, y en consecuencia no es pertinente.

## Problema 3 *(2 puntos)*

En un país el salario medio mensual en unidades corrientes de los trabajadores de un determinado sector productivo y los índices de precios de consumo a lo largo de los 4 últimos años fueron los siguientes:

| Año | Salario (€) | Índice de precios (100: 2010) |
|---|---|---|
| 2016 | 2043 | 108 |
| 2017 | 2125 | 110 |
| 2018 | 2400 | 125 |
| 2019 | 2600 | 140 |

Tareas a realizar:

(a) Calcular los índices de precios con base en 2016. Explicar cómo se hace.

(b) Expresar el salario en unidades monetarias constantes de 2016. Cómo se denomina a este proceso? Explicar cómo se hace.

(c) Calcular las variaciones anuales del salario en términos corrientes y constantes durante estos años, Comentar los resultados.

(d) Calcular la tasa media anual acumulativa de los salarios en términos nominales y reales.

---

(a)

Para calcular los índices con base en 2016, debemos igualar el índice correspondiente a 2016 a 100, y recalcular el resto proporcionalmente respecto de ese valor, con una simple regla de tres. Por ejemplo, para 2017: $110 -- > x // 108 -- > 100$, de modo que $x = (110/108) \times 100 = 101{,}85$.

| Año | Índice de precios (100: 2010) | Índice de precios (100: 2016) |
|---|---|---|
| 2016 | 108 | 100 |
| 2017 | 110 | $x = (110/108) \times 100 = 101{,}85$ |
| 2018 | 125 | $x = (125/108) \times 100 = 115{,}74$ |
| 2019 | 140 | $x = (140/108) \times 100 = 129{,}62$ |

(b)

Se denomina al proceso requerido *deflactación*. Para ello se dividen los salarios corrientes, nominales o en euros de cada ao, por el índice de precios correspondiente en tantos por uno (al que se denomina *deflactor*), en este caso con base 2016, porque se pide deflactar a precios de ese ao.

| Año | Salarios nominales (€) | Índice de precios (100: 2010) | Salarios constantes (€) (2016: 100) |
|---|---|---|---|
| 2016 | 2043 | 100 | 2043/1=2043 |
| 2017 | 2125 | 101.85 | 2125/1.0185=2086.4 |
| 2018 | 2400 | 115.74 | 2400/1.1574=2073.6 |
| 2019 | 2600 | 129.62 | 2600/1.2962=2005.8 |

(c)

Para calcular las variaciones anuales (mejor, interanuales) simplemente dividimos la magnitud de cada ao entre la del ao anterior:

| Año | Salarios nominales | Tasa de variación | Salarios contantes (2016: 100) | Tasa de variación |
|---|---|---|---|---|
| 2016 | 2043 | - | 2043 | - |
| 2017 | 2125 | $(2125/2043) \times 100 = 104{,}0$ | 2086.4 | $(2086{,}4/2043) \times 100 = 102{,}1$ |
| 2018 | 2400 | 112.9 | 2073.6 | 99.3 |
| 2019 | 2600 | 108.3 | 2005.8 | 96.7 |

Si bien los salarios nominales suben año a año, los salarios reales y por tanto el poder adquisitvo disminuye a partir del ao 2018, debido a la fuerte subida de la inflación a partir de ese año.

(d)

Para calcular la tasa media anual acumulativa, utilizamos la media geométrica. En relación a los salarios nominales:

$$\overline{\Delta s_n} = \left(\frac{2600}{2043}\right)^{1/3} = 1{,}083$$

Los salarios nominales han subido una media del 8.3 % anual.

Respecto a los salarios reales o en euros constantes:

$$\overline{\Delta s_r} = \left(\frac{2600}{2043}\right)^{1/3} = 0{,}994$$

Es decir, los salarios reales han disminuido de media un 0.6 % cada año.

## Problema 3 *(2 puntos)*

Un jugador participa en un juego de azar. Tira dos monedas al aire y por cada cara que sale gana 10 euros. No obstante, si salen dos cruces 50 euros.

Tareas a realizar, justificando el procedimiento:

(a) Calcular los valores que toma la variable aleatoria X: *ganancias/pérdidas del juego de azar*, así como su función de cuantía.

(b) Calcular la función de distribución de la anterior v.a. y representarla gráficamente.

(c) Calcular la ganancia media o esperada del juego y su varianza.

(d) Calcular $P[X = 10]; P[X \leq 10]; P[X = 15] : P[X \leq 15]$.

---

(a)

La variable aleatoria se concreta asignando a cada resultado del experimento un valor concreto (columnas 1, 2 y 3), y las función de cuantía asignando a cada valor de la v.a. una probabilidad (columnas 3 y 4).

Resultados del experimento: $\Omega : \{xx, cx, xc, cc\}$

| Resultado | Caras | x (v.a.) | P[X=x] |
|---|---|---|---|
| xx | 0 | -50 | 1/4=0.25 |
| cx,xc | 1 | 10 | 2/4=0.50 |
| cc | 2 | 20 | 1/4=0.25 |
| | | | 1 |

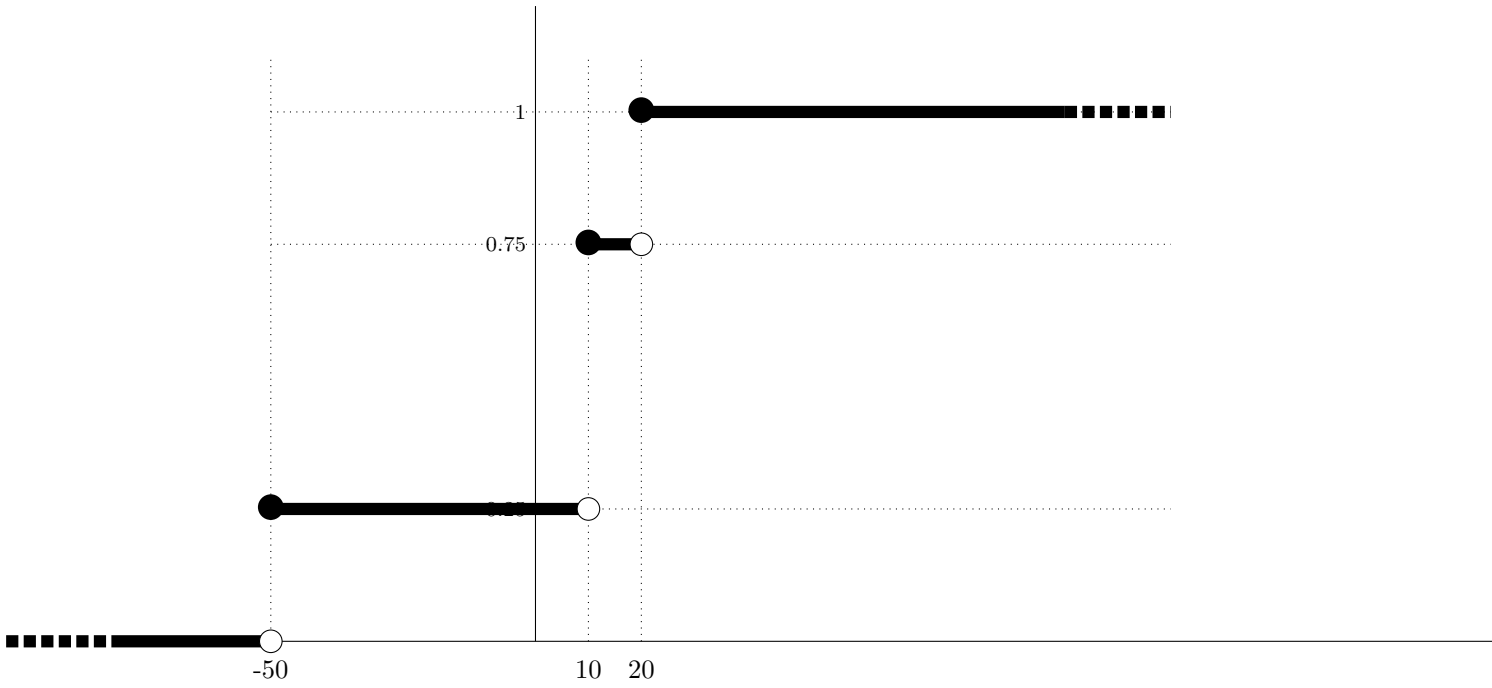
(b)

Acumulando la función de cuantía, obtengo la función de distribución en forma de tabla:

| x | F(x) |
|---|---|
| -50 | 0.25 |
| 10 | 0.75 |
| 20 | 1 |
| | 1 |

Matemáticamente: $F(x) = \begin{cases} 0 & x < -50 \\ 0{,}25 & -50 \leq x < 10 \\ 0{,}75 & 10 \leq x < 20 \\ 1 & x \geq 20 \end{cases}$

Gráficamente:



(c)

Calculemos el valor medio y la varianza de las ganancias. Para ello partimos de la función de cuantía:

| $x$ | $p(x)$ | $xp(x)$ | $x^2p(x)$ |
|---|---|---|---|
| -50 | 0.25 | -12.5 | 625 |
| 10 | 0.50 | 5 | 50 |
| 20 | 0.25 | 5 | 100 |
| | 1 | -2.5 | 775 |

$$E[X] = \sum xp(x) = -2{,}5$$

$$var[X] = \sum x^2p(x) - E[X]^2 = 775 - (-2{,}5)^2 = 768{,}75$$

El valor medio muestra que el juego es desfavorable a largo plazo.

(d)

Calculamos las probabilidades acumuladas con la función de distrbución y las simples con la función de cuantía:

- $P[X = 10] = 0{,}50$
- $P[X \leq 10] = 0{,}75$
- $P[X = 15] = 0$ (el valor 15 no pertenece al soporte de la distribución, es decir, no es posible una ganacia de 15 en una tirada)
- $P[X \leq 15] = 0{,}75$ (igual que $P[X \leq 10]$: evidente, ya que entre 10 y 20 no hay incremento de probabilidad)