

Corrección del examen parcial de Estadística administrativa I

Diciembre de 2006

1. El diagrama de barras y el histograma son representaciones gráficas se utilizan para obtener información visual sencilla acerca de la distribución de la variable. El diagrama de barras está pensado para representar variables ordinales o cardinales discretas y, en la práctica, se suele utilizar con variables que no tienen un número excesivo de modalidades en la muestra. El histograma, en cambio, es el gráfico más adecuado para representar variables continuas y, en la práctica, se utiliza con variables con un número elevado de modalidades.

Ambos gráficos se realizan sobre unos ejes de coordenadas. En el diagrama de barras se representan las **modalidades** de la variable en el eje de las equis y sobre cada modalidad se levanta una barra de altura proporcional a la frecuencia de esa modalidad. Todas las barras de este diagrama son de igual grosor, por lo que la frecuencia, representada por la altura, también es proporcional al área. En cambio en el histograma se representan los **datos agrupados en clases** en el eje de las equis, y sobre cada clase se levanta una barra de área proporcional a la frecuencia de esa clase. De esta forma, con ambos gráficos se puede valorar como es la distribución visualmente, observando el área (y/o la altura) de las distintas barras. Además, las barras del diagrama de barras están separadas, indicando de esta forma que los datos que se representan están aislados, mientras que en el histograma no hay espacio entre las barras, para indicar que los datos no están aislados, sino que se distribuyen de manera continua a lo largo del intervalo.

2. El **objetivo del Apartado a)** es calcular cuántos concejos presentan una renta disponible neta por habitante superior a los 12000 euros.

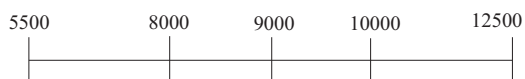
Planteamiento: el experimento consiste seleccionar concejos (individuos) y observar su renta disponible neta por habitante (variable). La población la constituirán entonces todos los concejos. La variable “renta disponible neta por habitante” es cardinal (porque los valores que toma son números), de razón (porque una renta de 0 euros significa no tener nada de renta) y continua (porque puede ser cualquier cantidad, por ejemplo, 8780 euros, 9024.45 euros, etc.). La muestra la constituyen todos los concejos asturianos, luego se trata de un censo. Como nos dan el rango en el que se encuentran los distintos concejos, vemos que tenemos datos agrupados.

Justificación: el objetivo lo podemos representar gráficamente como sigue:



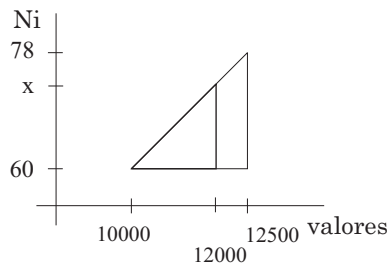
Es decir, por debajo de una renta de 12000 habrá un número a de concejos y por encima un número b . Nos piden calcular b , pero sabemos que $b = 78 - a$, y a partir de la tabla de frecuencias nos es más fácil calcular a . En definitiva, debemos calcular a , la posición absoluta del valor 12000 a partir de la columna de frecuencias absolutas acumuladas N_i de la tabla de frecuencias.

Cálculos: como tenemos datos agrupados, debemos localizar en primer lugar las clases de la tabla de frecuencias. Marcando los valores de renta que nos dan sobre una recta, vemos fácilmente que los intervalos son:



En la primera clase hay un 17.95 % de concejos, luego $n_1 = 78 \times 0,1795 = 14$ concejos. Nos dicen que hay 49 concejos por encima de 9000, luego por debajo habrá $N_2 = 78 - 35 = 43$ concejos. Así, sabemos que acumulados hasta 9000 hay 43, pero ya había 14 en el primer intervalo, luego en el segundo habrá $n_2 = 43 - 14 = 29$ concejos. También nos dicen que en el último intervalo hay $n_4 = 18$ concejos, y como en total suman 78, $n_3 = 78 - 14 - 29 - 18 = 17$ concejos. Tenemos así la columna de frecuencias absolutas n_i . Para completar la columna N_i , sólo falta acumular en el tercer y cuarto intervalo: $N_3 = 43 + 17 = 60$ y, por supuesto, $N_4 = 78$. En definitiva, la tabla de frecuencias es:

Clases	n_i	N_i	
[5500,8000]	14	14	Como 12000 está entre 10000 y 12500 (último intervalo), sabemos que el número de concejos acumulado hasta esa cifra estará ente 60 (N_i asociado a 10000) y 78 (N_i asociado a 12500). Para aproximar el valor que buscamos, podemos aplicar triángulos semejantes, suponiendo que las rentas están repartidos más o menos uniformemente en ese intervalo.
(8000,9000]	29	43	
(9000,10000]	17	60	
(10000,12500]	18	78	
Total	78	—	



Aplicando triángulos semejantes (base grande es a altura grande lo mismo que base pequeña es a altura pequeña), tenemos que:

$$\frac{12500-10000}{12000-10000} = \frac{78-60}{x-60}$$

En consecuencia $x - 60 = 18 \times 2000/2500$, es decir $x = 14,4 + 60 = 74,4$ concejos.

Interpretación: la posición absoluta de una renta de 12000 euros es de 74.4 concejos, que significa que hay aproximadamente unos 74.4 concejos con renta disponible por habitante de 12000 euros o menos y el resto, es decir, $b = 78 - 74,4 = 3,6$ tiene una renta de 12000 euros o más. En definitiva, podríamos decir que hay aproximadamente entre 3 y 4 concejos con renta disponible por habitante superior a los 12000 euros.

El **objetivo del Apartado b)** es encontrar la posición que ocupa Cudillero en relación con el resto de los concejos sabiendo que tiene una renta disponible neta por habitante de 9000 euros.

Planteamiento y justificación: El planteamiento es igual que en el apartado anterior, pero en esta ocasión nos piden una posición relativa, ya que debemos que decir el lugar que ocupa una renta de 9000 euros en comparación con el resto de los valores. Para hacer esto podemos bien tipificar (ya que no hay valores muy extremos y la media es representativa) o bien hallar el rango cuantil (ya que tenemos la tabla de frecuencias). Vamos a hallar, por ejemplo, el rango cuantil de 9000, es decir, diremos la posición que ocupa Cudillero a través del porcentaje de concejos con una renta igual o inferior a la suya.

Cálculos: para hallar el rango cuantil necesitamos la columna de frecuencias relativas acumuladas F_i , que se calcula dividiendo por $N = 78$ la columna de frecuencias absolutas acumuladas N_i obtenida en el apartado a). Como 9000 aparece en la tabla, directamente vemos que hay un 55.13 % de concejos con rentas iguales o menores a 9000.

Clases	F_i
[5500,8000]	0.1495
(8000,9000]	0.5513
(9000,10000]	0.7692
(10000,12500]	1
Total	—

Interpretación: la posición relativa de un concejo como Cudillero, con una renta de 9000 euros, es de 0.5513. Eso significa que un 55.13 % de los concejos tienen renta disponible por habitante de 9000 euros o menos y el resto, es decir, $100 - 55,13 = 44,87$ % tienen una renta de 9000 euros o más. En definitiva, podríamos decir que Cudillero ocupa una posición intermedia (ligeramente alta) en cuanto a renta disponible por habitante.

3. El objetivo del Apartado a) es indicar un valor alrededor del cual oscila el tiempo de desplazamiento diario de los encuestados.

Planteamiento: el experimento consiste seleccionar personas de cierta zona de Bélgica (individuos) y observar el tiempo de desplazamiento diario en coche a causa del trabajo en minutos (variable). La población la constituirán entonces todas las personas de esa zona. La variable “tiempo de desplazamiento diario en coche a causa del trabajo” es cardinal (porque los valores que toma son números), de razón (porque un tiempo de 0 minutos significa no estar nada en el coche) y continua (porque puede ser cualquier cantidad, por ejemplo, 28, 102.3 minutos, etc.). La muestra la constituyen todos los encuestados. No sabemos el tamaño muestral, y lo único que conocemos son los resultados el SPSS.

Justificación: como nos piden un valor alrededor del cual oscila la variable, tendremos que dar un valor central. De entre los datos que nos dan podríamos elegir entre la media, la mediana y la media recortada. Como la variable es cardinal, no hay ponderaciones y es de naturaleza aditiva y no multiplicativa, sabemos que la mejor medida es la media aritmética si no hay datos extremos. Si embargo, los resultados del SPSS indican que sí hay datos extremos. Esto se puede apreciar, por ejemplo, observando la gran diferencia que hay entre la media aritmética y la media recortada (60.5175 frente a 36.9001 minutos). Vemos que al quitar el 5 % de los valores más altos y más bajos, la media bajó casi hasta la mitad, lo que significa que hay valores bastante altos en comparación con el resto. Cuando hay datos extremos podemos decidir utilizar la media recortada (si la muestra es suficientemente grande) o la mediana. Podemos elegir, por ejemplo, la mediana.

Cálculos e interpretación: el SPSS nos informa directamente que la mediana es de 30 minutos, que significa que la mitad de los encuestados tardan en sus desplazamientos 30 minutos o menos y la otra mitad, tardan 30 minutos o más. En conclusión, podemos decir que el tiempo de desplazamiento diario oscila alrededor de 30 minutos.

El **objetivo del Apartado b)** es indicar la variabilidad conjunta en torno al centro.

Planteamiento y justificación: el planteamiento es igual que en el apartado a). Nos piden cuantificar la variabilidad (o dispersión) en torno a un centro. Nosotros conocemos medidas de dispersión para la media (pero no era buena medida de centro) y para la mediana (que si era una buena elección como representante central). En consecuencia, podremos indicar de

la variabilidad en torno a la mediana mediante el rango intercuartílico, que nos informa de lo dispersos o concentrados que se encuentran los tiempos moderados (quitando el 25 % de los más altos y el 25 % de los más bajos).

Cálculos e interpretación: el SPSS nos informa que el rango intercuartílico es de 23 minutos, lo que significa que, entre la gente que se desplaza en coche un tiempo moderado, hay una diferencia máxima de 23 minutos. Luego, podemos decir que la variabilidad conjunta en torno a los 30 minutos es de unos 23 minutos. Para evaluar si esa variabilidad es grande o pequeños, podríamos relativizar; como $23/30=0.7666$, diríamos que la variabilidad en relación a la mediana es del 76.66 %, es decir, el una variabilidad moderada/alta.

El **objetivo del Apartado c)** es explicar el valor de Curtosis e indicar si la distribución del tiempo de desplazamiento parece normal.

Planteamiento y justificación: el planteamiento es igual que en el apartado a). Nos piden explicar el valor de Curtosis, que mide el grado de apuntamiento de una variable en relación con una distribución de referencia que llamamos normal, es decir, si los valores centrales están más concentrados que los de la normal y los extremos son menos frecuentes o viceversa. La distribución normal es una abstracción de un tipo de distribuciones muy habituales en la práctica; es una distribución simétrica con muchos valores agrupados alrededor de la media y pocos bajos o altos.

Cálculos e interpretación: el SPSS nos informa que el coeficiente de curtosis es 65.31, pero dicho valor sólo se puede tener en cuenta si elegíamos la media como medida de tendencia central. De hecho sale un valor muy alto y sabemos que si no estuviera distorsionado por los valores extremos, no se alejaría más de unas pocas unidades del cero. Obviamente, la distribución de los tiempos de desplazamiento no se parece a la normal, ya que si fuese así, no habría valores muy extremos y curtosis sería aproximadamente cero. Además también observamos que no es una distribución simétrica, porque los percentiles 10 y el 90 no son en absoluto simétricos respecto al centro de la distribución.

El **objetivo del Apartado d)** verificar si hay gente que está más de 4 horas al volante para desplazarse al trabajo.

Planteamiento y justificación: el planteamiento es igual que en el apartado a). Para ver si hay gente en esas concisiones, necesitaríamos saber si hay algún valor que supere los $4 \times 60 = 240$ minutos en la muestra. Para ello podemos fijarnos en los percentiles y rangos, que son los únicos valores de los resultados que se refieren a individuos aislados.

Cálculos e interpretación: entre los percentiles y los rangos vemos que el percentil 10 es 0, luego había gente que no desplazaba en coche al trabajo. También vemos que el rango es 270, luego la diferencia entre el que menos tardó y el que más era de 270 minutos, es decir, de 4.5 horas. En consecuencia, vemos que sí hubo gente que tarda más de 4 horas, ya que al menos hay una persona que tarda 4.5 horas.

3. El objetivo del Apartado a) es verificar si el coste de producción total está muy repartido entre los distintos modelos o no.

Planteamiento: el experimento consiste seleccionar las unidades fabricadas de cada modelo (individuos) y observar el coste de producción (variable). La población la constituirán entonces todas las unidades fabricadas. La variable “coste de producción” es cardinal (porque los valores que toma son números), de razón (porque un coste de 0 euros significa no hay ningún coste) y continua (porque puede ser cualquier cantidad, por ejemplo, 3.2, 5 euros, etc.). La muestra la constituyen todos los coches producidos, es decir, tenemos una muestra de $12580+10530+5480+19500=48090$ unidades.

Justificación: para verificar si el coste de producción total está muy repartido entre los distintos modelos o no, tendremos que calcular el índice de Gini, que precisamente mide el grado de igualdad o desigualdad en el reparto del total de una variable entre los individuos de la muestra.

Cálculos: necesitamos construir la tabla de la desigualdad. En primer lugar, necesitamos saber cuál es el coste de producción total, y eso lo hacemos a través de la tabla de frecuencias (ordenando los costes de menor a mayor).

x_i	n_i	$x_i \times n_i$
4.20	19500	81900
6.40	10530	67392
7.80	12580	98124
12.60	5480	69048
Total	48090	316464

El coste total es la suma de la columna $x_i \times n_i$ (en la que se van calculando los costes parciales de cada modelo), por lo que vemos que es de 316464 euros. A continuación tenemos que ir calculando la proporción de coste acumulado que le va correspondiendo a cada modelo y comparándolo con la proporción coches manufacturados de cada modelo también acumulando.

La proporción de coste respecto al total del primer modelo es el cociente entre el coste de los coches producidos de ese modelo y el total, es decir, $q_1 = \frac{x_1 n_1}{x_T} = \frac{81900}{316464} = 0,2588$. Como al primer modelo se le dedicó una proporción de $F_1 = 19500/48090 = 0,4055$ de la producción total de coches, vemos que al 40.55 % de los coches más baratos se les dedicó el 25.88 % del coste total. Razonando de la misma manera con el segundo modelo más barato, tendríamos que la proporción

del coste respecto al total sería

$$q_1 = \frac{x_1 n_1 + x_2 n_2}{x_T} = \frac{81900 + 67392}{316464} = 0,4716,$$

mientras que la proporción de producción acumulada sería $F_2 = (19500 + 10530)/48090 = 0,6245$. Haciendo lo mismo para el resto de los modelos, podemos construir la tabla de desigualdad con las columnas F_i y q_i :

F_i	q_i	
4.20	0.4055	0.2588
6.40	0.6245	0.4716
7.80	0.8860	0.7818
12.60	1	1
Total	—	—

El índice de Gini cuantifica la desigualdad evaluando la diferencia entre los valores de q_i y los que se corresponden a la máxima desigualdad en términos relativos. En este caso sería:

$$I_G = \frac{\sum_{i=1}^{k-1} (F_i - q_i)}{\sum_{i=1}^{k-1} F_i} = \frac{(0,4055 - 0,2588) + (0,6245 - 0,4716) + (0,8860 - 0,7818)}{0,4055 + 0,6245 + 0,8860} = 0,4018$$

Interpretación: sabemos que si $I_G = 0$ significa que no hay desigualdad y si $I_G = 1$ la desigualdad es máxima. Entre 0 y 1, a medida que crece la desigualdad, crece el valor del índice de Gini. En este caso el índice de Gini es 0.4018, que indica una desigualdad considerable, luego podemos decir que el coste de producción total no está muy repartido entre los distintos modelos.

El **objetivo del Apartado b)** es determinar que parte de la inversión en producción se realizó con el modelo más caro.

Planteamiento, justificación y cálculos: el planteamiento es el mismo que el del apartado anterior. En dicho apartado, calculamos el coste parcial de cada modelo (columna $x_i \times n_i$) y el coste total (suma de esa columna). El modelo más caro es el último, así que debemos indicar qué parte del coste total, que era 316464, representa el último coste parcial, que era 69048. Para calcular cuánto es 69048 de 316464, debemos dividir: $69048/316464 = 0,2182$.

Interpretación: El coste del modelo más caro representa el 21.82 % del coste total.