

Corrección del examen parcial de Estadística administrativa I

Diciembre de 2005

2. La media recortada al 5 % es la media aritmética del 95 % central de los valores de la muestra. Es decir, se eliminan el 5 % de los valores extremos (el 2.5 % de los valores más altos y el 2.5 % de los más bajos) y se calcula la media de los que quedan. Al igual que la media aritmética, la media recortada es una medida de tendencia central para variables cardinales que tiene sus ventajas y sus inconvenientes. La ventaja es que al eliminar el 5 % los valores extremos, en caso de que haya valores muy alejados, éstos no influyen en su cálculo y en este sentido es mejor que la media. Sin embargo, eliminar individuos también supone un inconveniente, porque sabemos que las muestras son mejores cuanto más grandes sean y eliminar datos supone una pérdida de información. Si se dispone de muchos datos, eliminar alguno no es problema, pero si la muestra es pequeña, sí puede serlo. Además, es menos operativa matemáticamente que la media.

Si el tamaño de muestra es suficientemente grande, comparando los valores de la media aritmética y de la media recortada podemos saber si hay valores muy extremos que hacen que la media aritmética no sea buena. Si se diferencian poco, significa que el 5 % de los valores que quitamos no aportaban mucho en el cálculo de la media, pero si se diferencian mucho significa que tenían mucho peso y que la media puede estar distorsionada por ellos. En definitiva, utilizaremos este estadístico como medida de tendencia central únicamente si la muestra es suficientemente grande y hay valores muy extremos que hacen que la media esté muy desvirtuada.

2. El objetivo del Apartado a) es encontrar un *valor central* que resuma la distribución de los saldos de los clientes de esa sucursal, luego dependiendo de los datos podría ser la media, la mediana o la moda. Vamos a **plantear** el problema para ver que es lo más adecuado en este caso. El experimento consiste en ir eligiendo clientes (individuos) y observando el saldo de sus cuentas corrientes a 30 de Noviembre (variable). Por ejemplo, el cliente 1 puede tener un saldo de 3465 euros, el cliente 2 de 87 euros, etc. La población la constituirán entonces todos los clientes de esa sucursal y tenemos una muestra (que en este caso es un censo) de 166 clientes (la suma de los clientes en cada rango de saldos). La variable “saldo” es cardinal (porque los valores que toma son números), de razón (porque un saldo de 0 euros significa no tener nada en saldo) y continua (porque se puede tener cualquier cantidad, por ejemplo, 87 euros, 87.45 euros, etc.). Nos dan el número de clientes con saldos en diferentes rangos, es decir, las frecuencias absolutas de los saldos agrupados en clases:

Clases	n_i
Menos de 300	42
de 300 a 600	41
de 600 a 1200	34
de 1200 a 3000	22
Más de 3000	27
Total	166

Como la política de la sucursal impide los números rojos, la primera clase se pueden expresar como $[0, 300]$. Para evitar solapamientos, el resto podemos expresarlas como $(300, 600]$, $(600, 1200]$, $(1200, 3000]$ y $(3000, +\infty)$. Nos fijamos en que no conocemos el límite superior de la última clase y por eso utilizamos el símbolo ∞ . Sabemos que la mejor medida de tendencia central para variables cardinales suele ser la media, sin embargo, en este caso no podemos calcularla, porque no tenemos la marca de clase del último intervalo. En casos como estos, la mejor idea de centro la da la **mediana**.

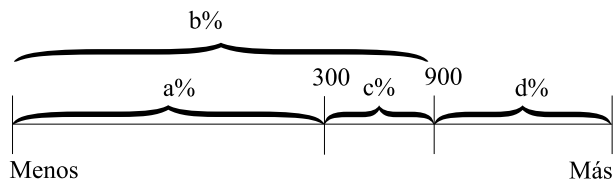
La mediana es un valor que deja la mitad de la muestra por debajo y la otra mitad por encima, es decir, es un valor que queda justo en el medio y por eso es buena para decir alrededor de qué cantidad se encuentran los saldos. Para calcularla construiremos la columna de frecuencias relativas acumuladas y buscaremos el punto en el que se acumula el 50 % de la muestra por debajo (exactamente si el 0.5 aparece en la tabla o aplicando triángulos semejantes en caso contrario):

Clases	n_i	F_i
$(0, 300]$	42	0.2530
$(300, 600]$	41	0.5000
$(600, 1200]$	34	0.7048
$(1200, 3000]$	22	0.8373
$(3000, +\infty)$	27	1
Total	166	—

La columna de frecuencias acumuladas representa la proporción de clientes con valor menor o igual que el extremo superior del intervalo. Por ejemplo, $F_1 = 42/166 = 0.2530$, que significa que el 25.3 % de los clientes tenían saldos de 300 euros o menos. Si añadimos ahora los de la siguiente clase, tendríamos $F_2 = 83/166 = 0.5000$, que significa que el 50 % de los clientes tenían saldos de 600 euros o menos, es decir, precisamente 600 euros es el valor de la mediana.

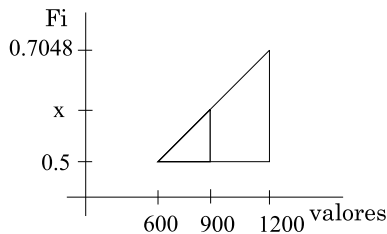
Podemos decir entonces que el saldo de los clientes de esa sucursal en cuentas corrientes oscilaba alrededor de 600 euros.

El objetivo del Apartado b) es calcular el porcentaje de clientes con saldos entre 300 y 900 euros. Si queremos utilizar la tabla de frecuencias para hacer el cálculo, debemos buscar siempre lo que se acumula por debajo de cada valor. El problema gráficamente sería:



Es decir, habrá el porcentaje c de clientes con saldos entre 300 y 900. Ese porcentaje lo podremos hallar como la diferencia entre lo que hay por debajo de 900 (que llamamos b) y lo que hay por debajo de 300 (que llamamos a). Para hallar a y b miramos la columna de frecuencias relativas acumuladas que hemos calculado antes.

Vemos que el valor de 300 euros es el extremo superior del primer intervalo, luego se corresponde con el primer F_i : sabemos que el 25.3 % de los clientes tenían saldos de 300 euros o menos, por lo que $a = 25,3\%$. Los 900 euros están en el intervalo $(600, 1200]$. Vemos que hasta 600 se lleva acumulado el 50 % y que en 1200 se lleva ya el 70.48 %. Para calcular cuanto se lleva aproximadamente hasta 900 podemos aplicar triángulos semejantes con esos valores (suponiendo que los saldos de los clientes entre 600 y 1200 están más o menos bien repartidos):



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

$$1200 - 600 = 600 \quad \text{—} \quad 0.7048 - 0.5 = 0.2048$$

$$900 - 600 = 300 \quad \text{—} \quad x - 0,5$$

En consecuencia $x - 0,5 = 300 \times 0,2048/600$, es decir $x = 0,1024 + 0,5 = 0,6024$. Podemos decir entonces que aproximadamente el 60.24 % de los clientes tienen saldos de 900 euros o menos. Es decir $b = 60,24\%$.

En conclusión, como el 60.24 % tiene 900 euros o menos y el 25.3 % tiene 300 euros o menos, tendremos que el $(60.24 - 25.3) = 34.94\%$ de los clientes aproximadamente contarían con saldos entre 300 y 900 euros.

5. El objetivo del Apartado a) es verificar si la media representa bien los datos de la muestra. En primer lugar, vamos a plantear el problema. El experimento considerado ha sido seleccionar ciudades españolas y observar el número de viviendas terminadas cierto año, luego la población la constituirán todas las ciudades y tenemos una muestra de 32. La variable “número de viviendas terminadas” es cardinal (porque los valores que toma son números), de razón (porque el 0 significa que no se terminó ninguna, es decir, expresa nulidad) y discreta (porque no se puede decir que se termina media vivienda, por ejemplo). No disponemos de los datos de todas las ciudades, sino de algunos resultados obtenidos con la ayuda del SPSS.

El número medio de viviendas terminadas representará bien los datos muestrales si éstos se parecen mucho a esa media, es decir, si la dispersión alrededor de la media es pequeña. La dispersión se puede medir con la desviación típica (o la aproximación que da el SPSS que se llama cuasi-desviación), pero es difícil saber si es un valor grande o pequeño, porque depende de las unidades que está midiendo, por eso es mejor relativizar respecto al valor medio y medir la dispersión en términos relativos con el coeficiente de variación (o su aproximación utilizando los cálculos del SPSS):

$$\text{Aproximación de CV} = \frac{\hat{S}_x}{\bar{x}} = \frac{230,2768}{482,4146} = 0,4773,$$

lo que significa que hay un 47,73 % de dispersión en relación a la media. Es un valor moderado, no muy bajo, por lo que no se puede decir que el valor medio represente perfectamente a la muestra, aunque tampoco podemos decir que no sea representativa.

El objetivo del Apartado b) es interpretar el coeficiente de curtosis. Como es negativo, la distribución es menos apuntada de lo normal, y eso significa que hay menos ciudades con un número moderado de viviendas terminadas de lo normal y hay más ciudades con muchas o pocas viviendas terminadas de lo que habría en una distribución normal. Como además es un valor bastante alejado de cero, significa que la diferencia con la normal es grande, es decir, los valores bajos, medios y altos están bastante más repartidos que lo que estarían en la distribución normal.

El objetivo del Apartado c) es determinar si hay valores extremos. En el gráfico de cajas vemos que el número normal de viviendas terminadas se mueve aproximadamente entre 200 y 900 (segmento inferior y superior) y no hay ningún valor que supere los límites de la normalidad (no hay ni círculos ni asteriscos), luego no hay ninguna ciudad en la muestra con un número desproporcionado de viviendas terminadas.

El objetivo del Apartado d) es determinar el rango de variación los 16 valores más moderados. Como hay un total de 32 ciudades, ese rango estará determinado precisamente por la caja del diagrama (50 % central), que parece estar aproximadamente entre unas 300 y 700 viviendas terminadas.

El objetivo del Apartado e) es saber si el valor de Lyon es más alto que el de Murcia en comparación con sus muestras. Para Lyon ya nos dan la puntuación tipificada (es decir, la comparación con el valor medio en términos relativos). Para Murcia

podemos hallar la aproximación de su puntuación tipificada con los datos de la tabla:

$$\text{Aproximación de puntuación tipificada de Murcia} = \frac{x_i - \bar{x}}{\hat{S}_x} = \frac{623 - 482,4146}{230,2768} = 0,6105,$$

lo que significa que Murcia está por encima del número medio de su muestra en términos relativos, pero menos de lo que está Lyon. Es decir, como la puntuación tipificada de Lyon (1.5385) es mayor que la de Murcia (0.6105), podemos decir que comparando con la muestra española y la francesa respectivamente, en Lyon se terminaron más viviendas que en Murcia.

4. El objetivo del Apartado a) es calcular *el consumo medio por habitante al día* en el conjunto de esos tres países. Vamos a **plantear** el problema para localizar los elementos que tenemos. Para calcular esa media el experimento debería consistir en seleccionar habitantes de esos tres países (individuos) y observar su consumo diario de cerveza. Por ejemplo, un irlandés podría consumir 1.2 litros, un danés 0.5 litros, etc. Nos fijamos en que disponemos de datos de un año, no de días, por lo que también podríamos considerar que nuestra experimento es observar el consumo anual (sabemos que si conocemos el consumo anual podemos hallar el consumo medio diario dividiendo por 365, así que eso no sería problema). La población la constituirán entonces todos los habitantes de esos tres países y tendríamos una muestra (que en este caso es un censo) de 91960800 personas (la suma de los habitantes de los tres países). La variable “consumo” es cardinal (porque los valores que toma son números), de razón (porque un consumo de 0 euros significa no consumir nada) y continua (porque se puede tener cualquier cantidad, por ejemplo, 0.5 litros, 0.33 litros, etc.). No disponemos de los datos de todos los habitantes por día, pero sabemos cuanto se consumió al año en cada país y el número de habitantes de cada país, así que podemos hacer las operaciones con esos números. En primer lugar, es sencillo calcular el consumo medio por habitante durante todo el año, porque sabemos que la media sería el consumo conjunto anual de todos los habitantes de los 3 países dividido por el número de habitantes de esos países, es decir,:

$$\text{Consumo medio al año} = \frac{\text{suma de los consumos anuales}}{\text{número total de habitantes}} = \frac{555000000 + 712000000 + 11225000000}{91960800} = 135,8405$$

Vemos entonces que el consumo medio de cerveza por habitante el año pasado fue de 135.8405 litros. Como el año tiene 365 días, para saber cuál fue el consumo medio diario, solo tenemos que dividir ese consumo medio anual por 365 días:

$$\text{Consumo medio al día} = \frac{135.8405}{365} = 0,3722$$

Se puede decir entonces que el consumo medio diario en esos tres países fue de 0.3722 litros, lo que significa que el consumo de cerveza de cada persona osciló en torno a 0.3722 litros al día.

El **objetivo del Apartado b)** es calcular *el porcentaje de cerveza que consume el país de menor consumo medio*, por lo que, en primer lugar, necesitamos saber cuál es el país de menor consumo medio por persona. Como el consumo medio es el el consumo total entre el número de habitantes, tenemos que el consumo medio el año pasado en Irlanda sería $555000000/4024600 = 137,9019$ litros por persona, análogamente el cálculo para Dinamarca daría 131.9105 litros por persona y en Alemania sería 135.9968 litros por persona, luego el país de menor consumo medio es Dinamarca.

La proporción de cerveza que se consume en Dinamarca es el consumo de ese país respecto del consumo total, es decir, $712000000/(555000000+712000000+11225000000)=0.057$, lo que significa que en el país de menor consumo medio (Dinamarca) se bebe el 5.7% del total de cerveza consumida en esos tres países.