Corrección del examen final de Estadística administrativa I

Febrero de 2006

1. Ambas medidas cuantifican la variabilidad o dispersión de los valores de la muestra en términos absolutos, pero de forma muy distinta. La **desviación típica** S_X mide lo alejados o próximos que están los valores de la muestra de la media aritmética en conjunto, por lo que también sirve para cuantificar lo representativo que resulta ese valor como medida resumen de todos los datos de la muestra (será muy representativa si todos los valores están cerca, es decir, si la desviación típica es "pequeña"), aunque al tener unidades es difícil de valorar. Matemáticamente se calcula a partir de la varianza S_X^2 como sigue:

$$S_X^2 = \frac{\text{suma de las distancias de los valores a la media al cuadrado}}{\text{número de individuos}} = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N} \text{ y } S_X = \sqrt{S_X^2}$$

Como vemos se calcula la media de las distancias al cuadrado de los valores de la muestra a la media aritmética (luego se hace la raíz cuadrada para que las unidades sean las mismas que las de la variable) . Si hay poca dispersión, esas distancias serán pequeñas (los valores se parecerán mucho a la media aritmética) y al contrario, si hay mucha dispersión, esas distancias serán grandes (los valores se parecerán poco a la media aritmética). Se utilizan las distancias al cuadrado y no el valor absoluto, por las ventajas matemáticas que presenta. La varianza también se puede calcular como $S_X^2 = \overline{x^2} - \overline{x}^2$ y esta segunda fórmula suele resultar más cómoda.

La **amplitud intercuartil** es la diferencia entre el tercer cuartil y el primero $C_3 - C_1$, (la anchura de la caja del gráfico de cajas) o en otras palabra, el rango en que se mueven "valores moderados" (eliminando el 25 % de los valores más bajos y el 25 % de los valores más altos). Este valor mide también la dispersión (pero no en relación a la media aritmética), ya que si es muy pequeño, significa que los valores moderados están muy próximos entre sí, y si es muy grande significa que hay mucha diferencia de unos a otros. Cuando utilizamos como medida de centro la mediana en lugar de la media, no podemos utilizar la desviación típica como medida de dispersión, pero sí podríamos utilizar la amplitud intercuartil.

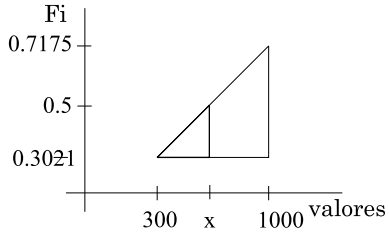
2. El **objetivo)** es calcular, en primer lugar, el umbral de pobreza de esa localidad, que es el 60 % del valor de la mediana, por lo que tenemos que comenzar calculando la mediana. Vamos a identificar los elementos del problema. **Planteamiento**: el experimento consiste seleccionar habitantes de esa localidad (individuos) y observar los ingresos mensuales que tiene (variable). La población la constituirán entonces todos los habitantes de la localidad y tenemos una muestra (o censo, porque parece que coincide con toda la población) de $7504 + 10319 + 6577 + 439 = 24839$ personas. La variable es cardinal (porque los ingresos son números), de razón (porque 0 significa no tener ningún ingreso, es decir, nulidad) y continua (porque, en principio, puede ser cualquier valor: 1234, 1234,6, etc.).

La **mediana** es el valor que deja la mitad de la muestra por debajo y la otra mitad por encima. Como tenemos datos agrupados 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]	7504	0.3021
(300, 1000]	10319	0.7175
(1000, 2000]	6577	0.9823
(2000,10000]	439	1
Total	24839	–

La columna de frecuencias acumuladas representa la proporción de habitantes con ingresos menores o iguales que el extremo superior del intervalo. Por ejemplo, F_1=7504/24839=0.3021, que significa que el 30.21 % de los habitantes de esa localidad disponen de ingresos de 300 euros o menos. Si añadimos ahora los de la siguiente clase, tendríamos $F_2 = (7504 + 10319)/166 = 0{,}7175$, que significa que el 71.75 % de los habitantes de esa localidad disponen de ingresos de 1000 euros o menos, por lo que la mediana tiene que estar entre 300 y 1000.

Para aproximar el valor de la mediana, lo más habitual es habitual es aplicar triángulos semejantes, porque es más exacto si los ingresos de los habitantes están repartidos más o menos de manera uniforme en ese intervalo (que es lo que más probable). La idea está en suponer que la F_i se va acumulando paulatinamente desde 0.3021 hasta 0.7175, es decir, siguiendo la pendiente del triángulo grande del gráfico, así que se trata de ver cuanto se lleva acumulado hasta 0.5, es decir, lo que corresponde a la pendiente del triángulo pequeño:



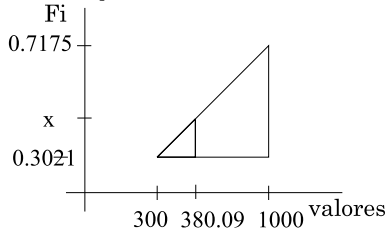
Siguiendo la regla de triángulos semejantes (base grande es a altura grande lo mismo que base pequeña es a altura pequeña), tenemos la siguiente regla de tres:

base grande=1000-300	———	altura grande=0.7175-0.3021
base pequeña=x-300	———	altura pequeña=0.5-0.3021

Es decir, $x-300 = (0{,}5-0{,}3021)\times(1000-300)/(0{,}7175-0{,}3021)$, por lo que $x = 633{,}4858$ euros. En conclusión, podemos decir que la mitad de los habitantes tienen unos ingresos de *aproximadamente* 633,4858 euros o menos y el 50 % restante está por encima de ese valor.

Como el **umbral de pobreza** es el 60 % de ese valor, sería $633{,}4858 \times 0{,}6$=380.0915. Se consideraría entonces que una persona es pobre si vive con ingresos inferiores a los 380,0915 euros aproximadamente.

Para saber el **porcentaje de población que vive por debajo de ese umbral**, tendríamos que calcular el porcentaje acumulado hasta los 380,0915 euros. Como tenemos datos agrupados, de nuevo tendríamos que aplicar triángulos semejantes. Los ingresos de 380,0915 euros se encuentran en el intervalo $(300, 1000]$. Como hasta 300 se acumula el 30.21 % de la población y en 1000 euros se lleva acumulado el 71.75, a 380,0915 le corresponderá un porcentaje entre esos valores. Aproximamos el porcentaje que buscamos como antes:



base grande=1000-300	———	altura grande=0.7175-0.3021
base pequeña=380.0915-300	———	altura pequeña=x-0.3021

Por lo que $x = 0{,}3496$, lo que significa que el 34.96 % de la población vive por debajo del umbral de la pobreza.

3. El **objetivo del apartado a)** es calcular los índices de calidad conjuntos base 2003 de cada año. Tenemos que comenzar planteando el problema para localizar la información que nos dan. **Planteamiento**: tenemos 3 bienes, que son los grupos de asignaturas. La importancia de cada bien es de 10, 8 y 7 respectivamente. La magnitud es la calidad. Los períodos son 2003 (que actuará como base), 2004 y 2005. Nos dicen que en el grupo 1 el índice de calidad disminuyó un 4 % en 2004 respecto a 2003, es decir, el índice estará 0.04 puntos por debajo de 1: $I_{03}^{04}(1) = 1 - 0{,}04 = 0{,}96$ puntos. En el grupo 2 el índice de calidad de 2004 respecto a 2003 aumentó un 2.5 %, es decir, el índice estará 0.025 puntos por encima de 1: $I_{03}^{04}(2) = 1 + 0{,}025 = 1{,}025$ puntos. En el grupo 3 ese índice se sitúa en 101.5 puntos porcentuales: $I_{03}^{04}(3) = 1{,}015$ puntos. Además, nos dicen que los índice de 2005 base 2004 son: $I_{04}^{05}(1) = 1{,}05$, $I_{04}^{05}(2) = 1{,}03$ y $I_{04}^{05}(3) = 0{,}995$. En conclusión, tenemos índices simples y ponderaciones (porque los grupos tienen distinta importancia), luego tendremos que calcular el índice conjunto media ponderada:

$$I_{03}^{04} = \frac{I_{03}^{04}(1)w_1 + I_{03}^{04}(2)w_2 + I_{03}^{04}(3)w_3}{w_1 + w_2 + w_3} = \frac{0{,}96 \times 10 + 1{,}025 \times 8 + 1{,}015 \times 7}{10 + 8 + 7} = 0{,}9962,$$

lo que significa que la calidad de la docencia en esos tres grupos de asignaturas bajó en media un $(1-0{,}9962)\times 100 = 0{,}38$ % en el año 2004 respecto a 2003 teniendo en cuenta la importancia de cada grupo.

Para hallar el índice del año 2005 base 2003 nos fijamos en que tenemos los índices simples de cada grupo del año 2005 base 2004, pero como también conocemos los índices de 2003 a 2004, podemos hacer un **enlace** para aproximar su valor en base 2003 (se podría hacer también el enlace con el índice conjunto): $I_{03}^{05}(1) = I_{03}^{04}(1) \times I_{03}^{04}(1) = 0{,}96 \times 1{,}05 = 1{,}008$, $I_{03}^{05}(2) = 1{,}025 \times 1{,}03 = 1{,}0558$, $I_{03}^{05}(3) = 1{,}015 \times 0{,}995 = 1{,}0099$. El índice conjunto sería entonces:

$$I_{03}^{04} = \frac{1{,}008 \times 10 + 1{,}0558 \times 8 + 1{,}0099 \times 7}{10 + 8 + 7} = 1{,}0238,$$

lo que significa que la calidad de la docencia en esos tres grupos de asignaturas subió en media un 2,383 % en el año 2005 respecto a 2003 teniendo en cuenta la importancia de cada grupo de asignaturas.

El **objetivo del apartado b)** el calcular la variación del índice conjunto base 2003 al pasar de 2004 a 2005 a causa de las asignaturas de del grupo 1. Podríamos contestar tanto en términos absolutos (repercusión absoluta) como en términos relativos (repercusión en porcentaje). Nos fijamos en que no podríamos contestar con la participación, porque *no* nos preguntan por la *parte* que se debe al grupo 1, *sino* por la *cantidad* en sí. Como

lo más habitual es medir dicha variación con la repercusión en porcentaje, es lo que haremos en este caso. Para hallar la repercusión en porcentaje hay que hallar primero la absoluta:

$$R_{03}^{04\to 05}(1) = \frac{(I_{03}^{05}(1) - I_{03}^{04}(1))w_1}{w_1 + w_2 + w_3} = \frac{(1{,}008 - 0{,}96) \times 10}{10 + 8 + 7} = 0{,}0192 \text{ puntos},$$

es decir, el índice conjunto de calidad base 2003 aumentó 0.0192 puntos al pasar de 2004 a 2005 a causa de las asignaturas del grupo 1. Para hallar la repercusión en porcentaje tenemos que dividir ese valor por el valor inicial del índice:

$$\%R_{03}^{04\to 05}(1) = \frac{R_{03}^{04\to 05}(1)}{I_{03}^{04}} = \frac{0{,}0192}{0{,}9962} = 0{,}0193,$$

lo que significa que los 0.0192 puntos representan un 1.93 % del valor inicial del índice, es decir, el índice conjunto de calidad base 2003 aumentó su valor un 1.93 % a causa de las asignaturas del grupo 1 al pasar de 2004 a 2005.
4. El **objetivo del apartado a)** se refiere a la regresión. Vamos a identificar en primer lugar todos los elementos del problema. **Planteamiento**: el experimento consiste en observar el número de vallas alquiladas (variable 1) y las ventas de refresco, en miles de litros, (variable 2) que se vende en cada ciudad (individuo). La población la constituirán entonces todas las ciudades de tamaño moderado de Europa y tenemos una muestra de tamaño desconocido. Las variables son cardinales (porque son respectivamente *número* de vallas y litros) y de razón (porque 0 significa no tener ninguna valla o no vender nada respectivamente, es decir, nulidad). El número de vallas alquiladas es discreta, porque no se puede alquilar 1 valla y media y las ventas es continua, porque pueden venderse 8953 litros o 8769.2 litros, etc.

En el apartado a) nos piden que expliquemos el mejor modelo de regresión. Para saber cuál es el mejor modelo, debemos mirar el que tiene mayor *coeficiente de determinación* (R^2 o Rsq), ya que ese valor mide la proporción de variación de Y que se explica por la curva correspondiente para ese modelo. Cuanto mayor sea R^2 mayor será esa proporción, es decir, más cerca estarán los datos de la curva correspondiente. En este caso, el mayor R^2 es el del modelo logarítmico Y=b0+(b1*ln(t)), con un valor de 0.87. Como queremos determinar las ventas *en función* del número de vallas que se alquilan, la *variable independiente* X es el número de vallas alquiladas (es la que controlaremos) y la *variable dependiente* Y son las ventas de refresco (es la que podremos predecir). La *fórmula de la regresión* sería entonces $y* = -900{,}07 + 534{,}661ln(X)$, eso *significa* que, de entre los modelos planteados, esta es la curva que mejor se aproxima a los datos. Como es la mejor fórmula para aproximar Y a partir de X, la regresión se puede *utilizar* para predecir los valores de las ventas en una ciudad si sabemos cuantas vallas se han alquilado en esa ciudad. La *fiabilidad* de la regresión para hacer predicciones se mide a través de su coeficiente de determinación. Es este caso $R^2 = 0{,}87$, lo que significa que el 87 % de la variabilidad de las ventas queda determinada por el número de vallas alquiladas a partir del modelo logarítmico. Como está bastante cerca del 100 %, eso significa que los datos se ajustan bastante bien a la curva, por lo que las predicciones serán bastante fiables.

El **objetivo del apartado b)** es calcular valores de Y a partir de valores de X. Para ello debemos utilizar el mejor modelo de regresión. Ya sabemos del apartado anterior que, en este caso, el mejor modelo es $y* = -900{,}07 + 534{,}661ln(X)$. Utilizando esa fórmula conseguiremos una *predicción*, no un valor exacto (nunca podemos asegurar que va a ser una cantidad exacta, siempre puede haber algún error) aunque sí bastante fiable, ya que el $R^2 = 0{,}87$ indica que los datos de la muestra se ajustan bastante bien a ese modelo. En primer lugar debemos asegurarnos que el número de vallas que nos dan están entre el mínimo y el máximo de la muestra, porque la regresión sirve para interpolar, pero no podemos decir si la curva es válida fuera del rango de la muestra (además las ciudades deberían ser del tipo de las de la muestra: de tamaño moderado y europeas). Nos dicen que en la muestra el número de vallas alquiladas variaba entre 10 y 150, luego podríamos hacer la predicción en el primer caso (valor de 100 vallas), pero no en el segundo (valor de 200 vallas). Sólo tenemos que sustituir X por 100 en la fórmula anterior: $y* = -900{,}07 + 534{,}661ln(100) = 1562{,}1349$ miles de litros (recordamos que la Y en la muestra venía medida en miles de litros). En conclusión, si se alquilan 100 vallas, pronosticamos unas ventas de 1562134.9 litros en los 6 meses (recordamos que esta aproximación es fiable, es decir, estará cerca del valor real, pero no es exacta).