

Corrección del primer parcial de Estadística administrativa I

Diciembre de 2008

1. Las medidas de dispersión están encaminadas a cuantificar lo próximos o alejados que están los datos de la muestra de un punto central. Estas medidas indicarán por un lado el grado de variabilidad que hay en la muestra y, por otro, la representatividad de dicho punto central, ya que si obtenemos un valor pequeño, eso significará que los valores se concentran entorno a ese centro (por lo que habrá poca variabilidad y el centro representará bien a todos) y viceversa.

Cuando el punto central elegido es la media se suele utilizar la desviación típica como medida de variación absoluta y el coeficiente de variación como medida de variación relativa, mientras que cuando el punto central es la mediana se suele utilizar el rango intercuartílico.

La desviación típica es la raíz cuadrada de la varianza, que a su vez es el promedio de la distancia al cuadrado de los valores de la muestra a la media. La varianza también es una medida de dispersión, pero viene medida en las unidades de la variable al cuadrado y, por lo tanto, no se puede interpretar fácilmente. Las unidades de la desviación típica son las mismas que las de la variable, por lo que es más sencilla de interpretar. Sin embargo, como depende de las magnitudes no se puede utilizar para establecer comparaciones ni valoraciones (como indicar si es 'grande' o 'pequeña') en general.

El coeficiente de variación relativiza el valor de la desviación en relación con la media. Al ser adimensional puede utilizarse para comparar/valorar la dispersión en distintas muestras. Sin embargo, por su cálculo cobra pleno sentido sólo cuando la variable es positiva y su media está alejada de 0.

El recorrido intercuartil indica la distancia máxima que hay entre los individuos de valores moderados (50 % central). Al igual que la desviación típica sus unidades son las mismas que las de la variable y no se puede utilizar para establecer comparaciones ni valoraciones en general.

2. El **objetivo del Apartado a)** es explicar el significado del índice de Gini.

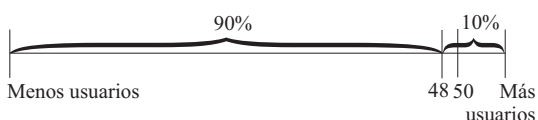
Planteamiento: el experimento consiste seleccionar meses (individuos) y observar el número de usuarios de un servicio de ayuda público ese mes (variable). La población la constituirán entonces todos los meses. La variable 'número de usuarios de un servicio de ayuda público cada mes' es cardinal (porque los valores que toma son números), de razón (porque 0 usuarios significa ausencia de usuarios) y discreta (porque puede ser 20, 21, pero no 20.5), aunque podría actuar como continua dependiendo del número de valores distintos que tome. La muestra es de tamaño desconocido. Nos dan una tabla con algunos descriptivos.

Método, justificación y cálculos: nos piden interpretar el índice de Gini (I_G), que es una medida adimensional de la desigualdad, y ya nos lo dan calculado. El índice de Gini varía entre 0 y 1 y se puede interpretar en términos de porcentaje. 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.

Conclusión: El índice de Gini indica una desigualdad del 31 %, por lo que el número de usuarios total se reparte de forma moderadamente desigual entre los distintos meses (hay meses en los que hay más gente y meses en los que hay menos, pero la diferencia no llega a ser excesiva).

El **objetivo del Apartado b)** es determinar si 50 usuarios al mes resulta un número elevado.

Planteamiento, método y justificación: El planteamiento es igual que en el apartado anterior, pero en esta ocasión nos piden una posición relativa, ya que debemos determinar si 50 usuarios al mes es mucho o poco en relación con el número de usuarios del resto de los meses. Para hacer esto podríamos bien tipificar o bien hallar el rango cuantil. Como no tenemos datos para la tipificación, sólo podemos intentar calcular o aproximar el rango cuantil con la información de los percentiles. Necesitamos la proporción $a\%$ de meses con 50 usuarios o menos. Gráficamente:



Cálculos: Con los datos que nos dan, no podemos calcular de forma exacta lo que necesitamos. El número de usuarios más cercano a 50 que aparece es 48, que se corresponde con el percentil 90. Gráficamente:

por lo que se puede decir que más de un 90 % de los meses hubo 50 usuarios o menos y menos del 10 % de los meses hubo 50 usuarios o más.

Conclusión: el valor 50 ocupa una posición relativa muy alta; la mayor parte de los meses (más del 90 %) hay menos de 50 usuarios, luego podemos decir que sí resulta un número elevado.

El **objetivo del Apartado c)** es determinar alrededor de qué valor oscila el número de usuarios al mes

Planteamiento, método, justificación y cálculos: El planteamiento es igual que antes, pero ahora debemos dar un valor central (es decir, un valor alrededor del cual se mueva la variable). Con los datos que tenemos sólo podemos utilizar la mediana (percentil 50 = 34), a pesar de que la media fuese mejor en caso de que no hubiese datos extremos.

Conclusión: Como la mediana es 34 podemos decir que el número de usuarios por mes osciló alrededor de 34.

El **objetivo del Apartado d)** es determinar si la distribución del número de usuarios por mes se parece a la normal en cuanto a la simetría

Planteamiento, método, justificación y cálculos: El planteamiento es igual que antes. Ahora debemos centrarnos en la simetría. La distribución normal es una abstracción de un tipo de distribuciones muy habituales en la práctica, con muchos valores agrupados alrededor de la media y pocos bajos o altos, repartidos de igual forma por encima y por debajo de dicho valor, es decir, simétrica.

El coeficiente de asimetría nos dice si una distribución cualquiera está distribuida de la misma manera por encima que por debajo de la media y, en ese caso, toma el valor 0. Este coeficiente es adimensional y suele variar entre -2 y 2 (más raramente entre -4 y 4). Según los cálculos de la tabla es de 0.012

Conclusión: como el valor de la asimetría es positivo indica que hay (pocos) meses que tienden a tener un número de usuarios alto, mientras que los meses con menos usuarios tienden a estar más concentrados. De todas formas, como el valor de la asimetría está bastante cercano a 0, podemos decir que la distribución es bastante simétrica y, por lo tanto, parecida a la normal en este aspecto.

3. El objetivo del Apartado a) es comparar el índice de masa corporal entre hombres y mujeres.

Planteamiento: el experimento consiste seleccionar hombres y mujeres (individuos) y observar su índice de masa corporal (variable). Tenemos 2 poblaciones: por un lado todos los hombres y por otro lado todas las mujeres. La variable ‘índice de masa corporal de cada persona’ es cardinal (porque los valores que toma son números), de razón (porque 0 ausencia de masa) y continua (porque puede ser 23, 23.1, etc), además como nos dan rangos vemos que efectivamente tomaba muchas modalidades distintas. La muestra es de 40 hombres y 40 mujeres. Nos dan rangos de variación y frecuencias (datos agrupados).

Método y justificación: tenemos que comparar el índice de masa corporal entre hombres y mujeres. Esa característica varía de persona en persona, por lo que no se puede hacer una comparación general. Lo mejor es resumir primero la distribución del índice para hombres y para mujeres utilizando una medida de tendencia central (que nos diga alrededor de qué valor se mueve ese índice para los hombres y para las mujeres) y luego hacer la comparación. Como la variable es cardinal, la mejor medida de tendencia central si no hay datos extremos es la media, ya que tiene en cuenta la magnitud de todos los valores, por lo que resulta muy informativa. Veremos si hay datos extremos (muy alejados del resto y poco frecuentes) al construir la tabla de frecuencias. En caso de que los haya, optaremos por la mediana (aunque sólo tiene en cuenta el orden y resulta, por lo tanto, menos informativa). Como la variable es de razón la comparación la haremos mediante el cociente, porque es más sencillo de interpretar que la diferencia, ya que no depende de las magnitudes.

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



En la primera clase hay 5 mujeres y 2 hombres (frecuencia absoluta). Nos dicen que un 35 % de las mujeres y un 55 % de los hombres están por encima de 24, luego por debajo tendríamos una frecuencia relativa acumulada de $F_i = 0,65$ para las mujeres y $F_i = 0,45$ para los hombres.

Como hay un total de 40 mujeres y de 40 hombres, para el valor 24 tenemos acumuladas $N_i = 0,65 \times 40 = 26$ mujeres y $N_i = 0,45 \times 40 = 18$ hombres. Así que restando nos queda que con índice entre 20 y 24 hay $n_i = 26 - 5 = 21$ mujeres y $n_i = 18 - 2 = 16$ hombres. Además nos dicen que la frecuencia absoluta de la clase 28-32 es de 5 para las mujeres y 8 para los hombres. El resto de personas hasta sumar 40 tienen que estar en el intervalo de 24 a 28 (serían $40 - 5 - 21 - 5 = 9$ mujeres y $40 - 2 - 16 - 8 = 12$ hombres). En definitiva, la tabla de frecuencias es:

Clases	Mujeres n_i	Hombres n_i	x_i	Mujeres $x_i n_i$	Hombres $x_i n_i$
[18,20]	5	2	19	95	38
(20,24]	21	16	22	462	352
(24,28]	9	14	26	234	364
(28,32]	5	8	30	150	240
Total	40	40		941	994

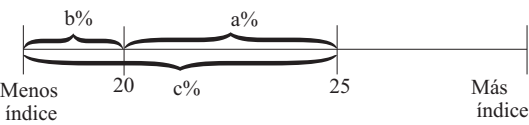
Vemos que no hay ningún dato excesivamente alejado del resto con frecuencia baja, luego calcularemos las medias aproximadas a partir de las marcas de clase y la columna $x_i n_i$. La media para las mujeres sería $941/40 = 23,525$ y para los hombres $942/40 = 24,85$. El cociente quedaría $23,525/24,85 = 0,9467$, es decir, el valor medio para las mujeres se sitúa un $(1 - 0,9467) \times 100 = 5,33\%$ por debajo del de los hombres.

Conclusión: El índice de las mujeres de la muestra oscila alrededor de 23.525 y el de los hombres alrededor de 24.85. Así, las mujeres presentan un índice de masa corporal un 5.33 % inferior al de los hombres en media.

El **objetivo del Apartado b)** es calcular el porcentaje de personas saludables en esa muestra.

Planteamiento: la única diferencia con el planteamiento de a) es que ahora se considera una solo población que involucra todos los hombres y mujeres. Por tanto tenemos ahora una muestra de 80 personas y la frecuencia absoluta de cada clase será la suma de las frecuencias absolutas de hombres y mujeres.

Método y justificación: Las personas saludables son las que tienen un índice entre 20 y 25, luego debemos calcular el porcentaje de personas en ese intervalo. Es decir:

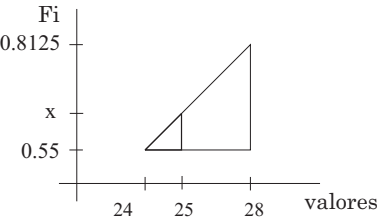


Nos preguntan por el porcentaje “a %” que hay entre 20 y 25, pero lo más sencillo es utilizar la tabla de frecuencias y calcular primero “b %” y “c %” con la columna F_i . Así a se calculará simplemente como la diferencia entre c y b .

Cálculos: La tabla de frecuencias de todas las personas con las frecuencias relativas acumuladas F_i es:

Clases	n_i	F_i
[18,20]	7	0.0875
(20,24]	37	0.55
(24,28]	21	0.8125
(28,32]	13	1
Total	40	–

El valor $b = 8,75\%$ se obtiene de manera exacta de la tabla, ya que es el porcentaje de gente con masa menor o igual que 20, es decir, la frecuencia relativa acumulada del primer intervalo. El valor c se correspondería con la frecuencia acumulada hasta 25, pero como dicho valor no es el extremo de ningún intervalo (si no que se encuentra en la clase 24-28), tendremos que utilizar triángulos semejantes para aproximar (suponiendo que las índices están repartidos más o menos uniformemente en ese intervalo).



Aplicando triángulos semejantes (base grande es a altura grande lo mismo que base pequeña es a altura pequeña), tenemos que:
 $24-28=4$ $0.8125-0.55=0.2625$
 $25 - 24 = 1$ $x - 0,55$
 En consecuencia $x - 0,55 = 0,2625/4$, es decir $x = 0,55 + 0,0656 = 0,6156$.

Aproximadamente el $c = 61,56\%$ de la muestra tiene un índice de 25 o menos. Así que $a = 61,56 - 8,75 = 52,81\%$.

Conclusión: aproximadamente el 52.81 % de la muestra tiene un índice de masa corporal comprendido entre 20 y 25 y son, por lo tanto, personas saludables.