CÁLCULO Y COMENTARIOS SOBRE ALGUNAS MEDIDAS DESCRIPTIVAS

I Medidas de localización

Aunque una distribución de frecuencias es ciertamente muy útil para tener una idea global del comportamiento de los datos, es generalmente necesario resumir los datos aún más, calculando algunas medidas descriptivas. Estas medidas son valores que se interpretan fácilmente y nos sirven para un análisis más profundo que el obtenido por medio de resúmenes gráficos y tabulares.

En esta sección calcularemos *medidas de localización*, es decir, medidas que buscan cierto lugar del conjunto de datos; cuando el lugar buscado es el centro de los datos les llamamos *medidas de tendencia central*, entre las que veremos: la media, la moda y la mediana.

La *media muestral* de un conjunto de n observaciones $x_1, x_2, ...x_n$ de una variable X, la denotaremos por $\bar{x}$ y la calcularemos mediante la fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

No existe una regla general acerca de cuántos decimales reportar en el resultado de este cálculo, pero no tiene sentido alejarse mucho del número de decimales que poseen los datos. Podemos tomar un decimal más que éstos. Nótese que la media sólo tiene sentido para datos a nivel de intervalo o de razón y que el valor de la media muestral puede variar de muestra a muestra.

La *mediana* de un conjunto de n observaciones ordenadas $x_1, x_2, ...x_n$ es el valor que divide el conjunto de datos en dos partes iguales. Podemos denotar a la mediana por $\tilde{x}$ (x tilde). Para encontrar la posición o lugar dónde buscar la mediana en un conjunto de n observaciones calcularemos:

$$\textit{Posición de la mediana} = (n+1)/2$$

Así, cuando n es impar, la posición de la mediana coincide con el lugar de uno de los datos. Si n es par, se localizará en medio de los dos datos centrales.

La *moda* de un conjunto de *n* observaciones $x_1, x_2, ...x_n$ es el valor que se repite con mayor frecuencia. La podemos denotar por $\hat{x}$. Mientras que la mediana se aplica a datos ordenados, la moda puede aplicarse a datos nominales.

Por ejemplo, si deseamos calcular la media, la moda y la mediana de los datos 12, 14, 15, 17, 18, 18 y 22, tendríamos:

La media es $\bar{x} = (12+14+...+22)/7 = 16.57$.
Como la posición de la mediana es (7+1)/2 = 4, la mediana es $\tilde{x} = 17$.
Y, finalmente, la moda sería el 18, esto es, $\hat{x} = 18$.

Existen otras medidas de localización para datos ordenados, como son los cuartiles, deciles y percentiles. Los *cuartiles* dividen en cuartos una distribución de frecuencias. Los denotaremos por q_1, q_2 y q_3. El segundo cuartil coincide con la mediana. Los *deciles* los denotaremos por $d_1, d_2, ...d_9$ y son valores que dividen una distribución de frecuencias en diez partes iguales. El quinto decil coincide con la mediana. Los *percentiles* se denotarán por $p_1, p_2, ...p_{99}$ y son valores que dividen una distribución de frecuencias en cien partes iguales.

Podemos encontrar fórmulas análogas a la de localización de la mediana, para localizar los otros cuartiles, los deciles y los percentiles:

Posición de $q_1 = (n+1)/4$
Posición de $q_3 = 3(n+1)/4$
Posición de $d_7 = 7(n+1)/10$
Posición de $p_{85} = 85(n+1)/100$

Hay que recordar que los cálculos anteriores nos dan la posición donde debemos buscar las medidas anteriores. Ya encontrada la posición, existen algunos criterios para asignar el valor correspondiente a la medida buscada. Por ejemplo, si la posición del percentil 85 es el lugar 23.42, algunos toman el dato que está en el lugar 23, otros el que se encuentra en el lugar 24 y otros el dato 23 más 0.42 veces la diferencia que existe entre el dato 24 y el dato 23, aunque esto último sólo tiene sentido cuando los datos se miden a nivel de intervalo. Al utilizar paquetería estadística se puede observar que puede haber diferencias en estos cálculos, y es debido a esa razón. Lo mismo se puede observar en el cálculo de cuartiles en las calculadoras.

Podemos también calcular las medidas de centralización cuando tenemos los datos agrupados en una tabla de frecuencias. Si denotamos por m_i la marca de clase del i-ésimo intervalo y f_i su frecuencia absoluta, entonces la media aritmética se calculará como:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

Donde n es el tamaño total de la muestra y k es el número de intervalos en la tabla.

Con respecto a la moda, aunque existen algunas fórmulas para calcularla en una tabla de frecuencias, podemos tomarla simplemente como la marca de clase del intervalo con mayor frecuencia. Una de las fórmulas más usadas es la siguiente, que toma en cuenta las frecuencias de las clases contiguas a la clase modal:

$$\hat{x} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c$$

donde

L es la frontera inferior de la clase modal,

Δ_1 = frecuencia de la clase modal – frecuencia de la clase anterior,

Δ_2 = frecuencia de la clase modal – frecuencia de la clase siguiente,

c = longitud del intervalo de clase.

Para calcular la mediana, supondremos que los datos se distribuyen en una forma continua. Así, la mediana es aquel dato que corresponde a la mitad de la frecuencia total, o sea *n/2*, es decir, que deja la mitad de frecuencias por arriba y la otra mitad por debajo. Para ejemplificar su cálculo, que básicamente es una interpolación, consideremos la tabla de frecuencias, que se muestra a continuación:

Fronteras de clase	Marcas de clase	Frecuencias absolutas	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
3.55-5.45	4.5	1	1/40	1	1/40
5.45-7.35	6.4	2	2/40	3	3/40
7.35-9.25	8.3	9	9/40	12	12/40
9.25-11.15	10.2	9	9/40	21	21/40
11.15-13.05	12.1	14	14/40	35	35/40
13.05-14.95	14.0	3	3/40	38	38/40
14.95-16.85	15.9	2	2/40	40	1

En esta tabla la frecuencia total n es de 40, de modo que buscaremos la mediana en el dato número $40/2 = 20$. Si nos fijamos en las frecuencias absolutas acumuladas, encontraremos que la mediana está en el cuarto intervalo, ya que hasta el tercero llevamos una frecuencia acumulada de 12. Usando interpolación lineal, la mediana será

$$\tilde{x} = 9.25 + (8/9)(11.15 - 9.25) = 10.93;$$

o sea

$$\tilde{x} = L + \frac{\left(\frac{n}{2} - F\right)}{f} c,$$

donde

L es la frontera inferior de la clase mediana,

n es el número de datos de la muestra,

F es la frecuencia acumulada antes de la clase mediana,

f es la frecuencia de la clase mediana,

c es la longitud del intervalo de clase.

Usando un procedimiento similar, se pueden calcular los otros cuartiles, los deciles y los percentiles.

La media aritmética y la moda de la tabla anterior serían:

$$\bar{x} = \left[(4.5)(1) + (6.4)(2) + ... + (15.9)(2)\right] / 40 = 10.67$$

$\hat{x} = 12.1$ o $\hat{x} = 11.74$,

Según se tome la marca de clase o se utilice la fórmula para obtener la moda.

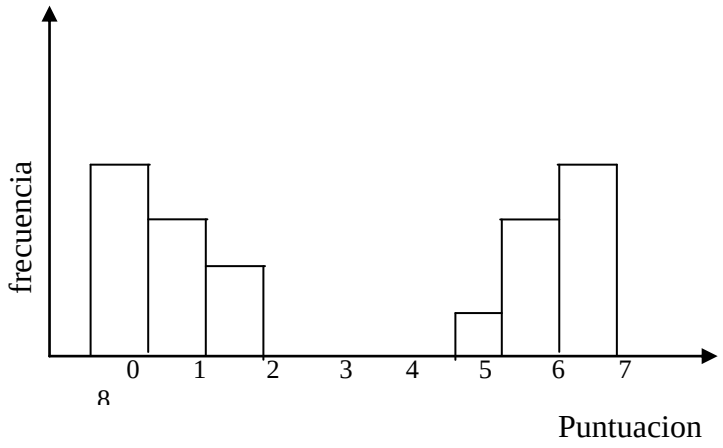
Elección de una medida de tendencia central o de localización

El cálculo de la moda, mediana o media aritmética es puramente mecánico y actualmente esto se hace con mayor rapidez en las computadoras e incluso en las calculadoras. Sin embargo, la elección entre estas tres medidas y su interpretación puede algunas veces requerir detenidas reflexiones. A continuación se presentan algunas consideraciones que deben estar presentes en mente cuando se esté haciendo frente a la elección:

1. En un grupo pequeño de datos la moda puede ser completamente inestable. Por ejemplo la moda del grupo (1,1,1,3,5,7,7,8) es 1; pero si uno de los unos se cambia por 0 y el otro por 2, la moda se convierte en 7.

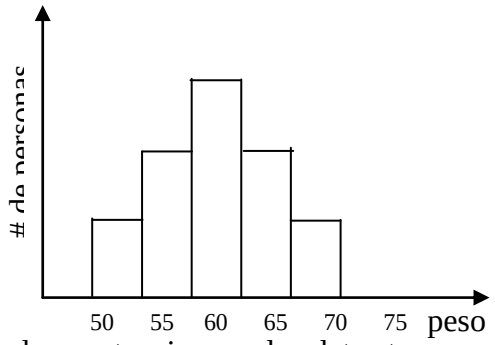
2. La media se ve influida por el valor de cada puntuación del grupo de datos. Si una puntuación cualquiera cambia por c unidades, se cambiaría en la misma dirección por c/n unidades. Por ejemplo, si 100 se suma a la tercera puntuación mayor en un grupo de 10, la media del grupo se aumentara en 10 unidades.
3. La mediana no se afecta por un cambio en el valor mayor o menor. Por ejemplo, en un grupo de 50 puntuaciones o datos la mediana no cambiaría si la puntuación mayor se triplica.

4. Algunos grupos de puntuaciones o datos simplemente no manifiestan tendencia central alguna en forma significativa, siendo a menudo engañoso calcular una medida de tendencia central. Esto es particularmente cierto para grupos de datos con más de una moda. Por ejemplo en la siguiente situación: Un investigador en desarrollo curricular, sostiene que se pueden construir pruebas de rendimiento compuestas por 8 ítems de elección múltiple que separan a los estudiantes entre los que han adquirido el concepto de suma de dos números y en los que no lo han adquirido. Los que lo adquirieron se representan con las puntuaciones 6,7,8, y los que no lo adquirieron se representan con puntuaciones de 0,1, y 2. Supongamos que un grupo de estudiantes da lugar a las puntuaciones que se presentan en el siguiente histograma de frecuencias que a continuación se presenta.

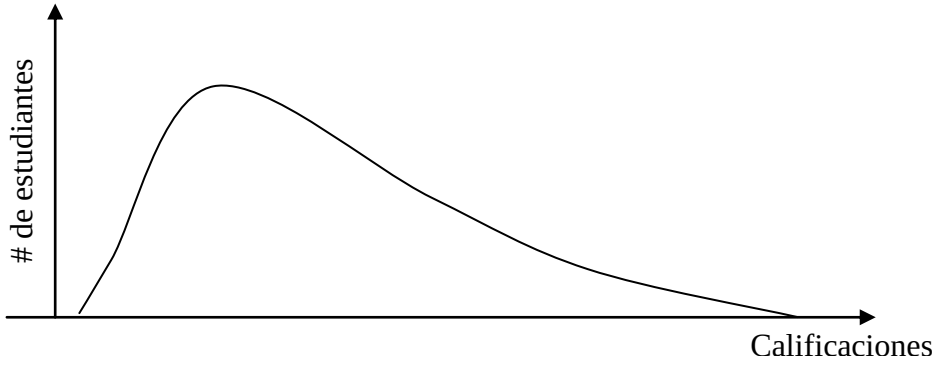


La media de las puntuaciones representadas, estaría en el rango de 3 a 5 a pesar de que nadie obtuvo ninguna de esas puntuaciones. La mediana del grupo esta aproximadamente en el mismo rango. En este caso ni la media aritmética ni la mediana representan adecuadamente a este grupo de puntuaciones o datos, tal vez la medida adecuada sea la moda, mas precisamente bimodal, ya que una moda sería 0 y la otra sería el 8.

5. La moda es posible localizarla tanto en variables cuantitativas, como cualitativas; la mediana también, si la variable cualitativa es de escala ordinal.
6. La medida de tendencia central en grupos de puntuaciones con valores extremos se mide probablemente mejor por la mediana, si puntuaciones o datos son unimodales. Como indicamos previamente, cada dato en un grupo influye en la media. Así, un valor extremo puede alejar a la media de un grupo de su valor inicial, de lo que generalmente se considera como la región central. Por ejemplo, si nueve personas tienen ingresos mensuales que fluctúan de $4500 a $ 5200 con un promedio de $4900 y el ingreso de una décima persona es de $20000, el ingreso promedio del grupo de las 10 personas es de $6410, Este valor no representa adecuadamente a ninguno de los grupos. La mediana sería en este caso preferible como medida de tendencia central.
7. En grupos unimodales de datos o puntuaciones simétricas la mediana, moda y media aritmética son iguales. Como se ilustra en la figura siguiente:



8. En el caso de que las puntuaciones o los datos tengan una marcada asimetría o sesgo como el que se ilustra en la siguiente figura, la moda será menor que mediana y esta a la vez, menor que media aritmética. En el caso de existir sesgo en la dirección contraria entonces la media aritmética será menor que la mediana y esta a su vez menor que la moda.



II Algunas Medidas de Dispersión

Puesto que esperamos que las características que medimos en la muestra reflejen de alguna manera las características de la población, mediremos la variabilidad en la muestra para entender la variabilidad que existe en la población. Como medidas de variabilidad estudiaremos el rango muestral, el rango intercuartílico, la varianza, la desviación estándar y el coeficiente de variación.

El *rango muestral* ya lo hemos calculado anteriormente al construir tablas de frecuencias, y es la diferencia entre el dato mayor y el menor. El *rango intercuartílico*, como su nombre lo indica, es la diferencia entre el tercer y el primer cuartil. Si lo denotamos por *RI*, tenemos que $RI = q_3 - q_1$.

La *varianza muestral* de un conjunto de n observaciones $x_1, x_2, ...x_n$ de una variable aleatoria X, se denota por s^2 y se calcula mediante la fórmula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Se utilizará como denominador $n-1$ en lugar de n, pues, como estimador de la varianza poblacional, cuando se divide entre $n-1$ tiene la propiedad de ser insesgado, es decir, de dar valores cuyo promedio es la varianza poblacional, como se verá en el capítulo cuatro.

Cuando el cálculo de la varianza muestral se hace en calculadora, sin utilizar las funciones estadísticas que muchas de éstas tienen, es más rápido y seguro utilizar cualquiera de las fórmulas siguientes, que fueron obtenidas simplificando la fórmula de la definición de varianza e implican un número menor de operaciones.

$$s^2 = \frac{\left[\sum_{i=1}^{n} x_i^2 - (1/n)\left(\sum_{i=1}^{n} x_i\right)^2\right]}{n-1} \quad \text{o bien} \quad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

La *desviación estándar* de un conjunto de datos es simplemente la raíz cuadrada positiva de la varianza. La denotaremos por *s*, y entonces

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

La varianza y la desviación estándar son medidas de variación absoluta y dependen de la escala de medición; sin embargo, hay ocasiones en que se necesita comparar la variación de diferentes conjuntos de datos y se requiere una medida de variación relativa, como el *coeficiente de variación*, en el que la desviación estándar se expresa como un porcentaje de la media. Lo calcularemos así:

$$\text{Coeficiente de variación} = \frac{s}{\bar{x}}(100)$$

Por ejemplo, calculemos el rango, el rango intercuartílico, la varianza y la desviación estándar del siguiente conjunto de datos: 12, 14, 16, 19, 19, 20 y 23.

$$x_{\max} = 23, \qquad x_{\min} = 12, \qquad q_1 = 14, \qquad q_3 = 20;$$

entonces

$$rango = 23 - 12 = 11, \qquad RI = 20 - 14 = 6,$$

$$\bar{x} = 17.6,$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - (1/n)\left(\sum_{i=1}^{n} x_i\right)^2}{n-1} = \frac{\left(12^2 + 14^2 + ... + 23^2\right) - \left(12 + ... + 23\right)^2 / 7}{6} = \frac{\left(2247 - 15129/7\right)}{6} = 14.28$$

y, así, la desviación estándar es la raíz cuadrada positiva de 14.28, esto es, $s = 3.8$. Por otra parte, el coeficiente de variación sería:

$$\text{Coeficiente de variación} = \frac{s}{\bar{x}}(100) = \frac{3.8}{17.6}(100) = 21.6\%.$$

Como ya vimos en la sección anterior, podemos calcular la media y los cuartiles en una tabla de frecuencias. De igual manera podemos calcular la varianza y, por lo tanto, la desviación estándar y el coeficiente de variación. La fórmula que utilizaremos para la varianza en una tabla de frecuencias es:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k} m_i^2 f_i - \frac{1}{n}\left(\sum_{i=1}^{k} m_i f_i\right)^2\right)$$

donde, como se dijo anteriormente, m_i y f_i son la marca de clase y la frecuencia absoluta del i-ésimo intervalo, *k* es el número de intervalos de clase y n es el número de datos.

Así como contamos con medidas de localización y de dispersión, que nos describen ciertas características de una distribución de frecuencias, existen otras medidas que nos pueden ayudar a distinguir cuestiones como simetría o grado de apuntamiento de una distribución.

Una distribución que no es simétrica, sino que se extiende más hacia uno de los extremos o colas, se denomina *sesgada*. Si la cola más larga se extiende a la derecha, se dice que la distribución está *sesgada a la derecha*, mientras que si la cola más larga se extiende a la izquierda, se dice que la distribución está *sesgada a la izquierda*. El *sesgo* se puede calcular utilizando los momentos de una variable aleatoria o de una distribución,

pero podemos calcular una medida alternativa de sesgo que emplea conceptos que ya hemos manejado. Esta medida se calcula como:

$$Sesgo = 3(\bar{x} - \tilde{x}) / s,$$

se llama *segundo coeficiente de sesgo de Pearson* y toma valores entre –1 y 1. Valores negativos indicarán un sesgo a la izquierda y valores positivos, un sesgo a la derecha.

Otra característica de la forma de una distribución se llama *curtosis* y nos indica el grado de apuntamiento de la distribución. Si la distribución es parecida a una distribución normal, que tiene una forma conocida también como "campana de Gauss", se le llama *mesocúrtica*. Si la distribución presenta un apuntamiento más alto que el de una distribución normal, se le llama *leptocúrtica* y en el caso de presentar menos apuntamiento que la distribución normal, se le llama *platicúrtica.* Al igual que el sesgo, la curtosis se puede calcular usando momentos de una variable aleatoria, pero por ahora usaremos una fórmula que involucra conceptos ya utilizados. Así, tenemos que el *coeficiente de curtosis percentílico* es:

$$K = \frac{q_3 - q_1}{2(p_{90} - p_{10})}.$$

Con este coeficiente de curtosis, cuyos valores se encuentran entre 0 y 1, una distribución es mesocúrtica si $K = 0.263$, leptocúrtica si $K < 0.263$ y platicúrtica si $K > 0.263$.

Para muestras provenientes de una distribución normal, el sesgo y la curtosis no tomarán necesariamente el mismo valor, sino que fluctuarán debido a la variación muestral.