Estudio descriptivo de una colección de datos

Cuando se ha recolectado la información correspondiente al fenómeno que se está investigando, se cuenta con una colección de datos individuales, la cual constituye la materia prima para el investigador. Comúnmente, este conjunto de datos es bastante grande y por ende es muy difícil obtener algunas conclusiones que sean de utilidad para el estudio. Por tal razón se hace necesario utilizar los métodos estadísticos descriptivos tanto para resumir y presentar convenientemente los datos, como también para conseguir algunos indicadores numéricos que sean de utilidad para la interpretación de los aspectos más importantes y de interés de los datos.

Organización de datos cualitativos

La manera de condensar o agrupar los datos cualitativos es muy intuitiva. Sólo es necesario un conteo de las distintas modalidades que presenta la variable en cuestión, lo que se conoce como frecuencia:

VARIABLE			
Modalidad 1	**. . .**	**Modalidad k**	**Total**
f_1	. . .	f_k	n

Tabla de doble entrada o tabla de contingencia

También se pueden organizar dos variables en una tabla. Este tipo de organización de datos se conoce como *tabla de doble entrada o tabla de contingencia*:

		VARIABLE A a_1	a_2 …	a_i	**TOTALES**
VARIABLE B	b_1				
	b_2				
	⋮				
	b_j				
TOTALES					

Nota:
También pueden organizarse en una misma tabla tres o más variables.

Organización de datos cuantitativos

Cuando se agrupan datos cuantitativos generalmente el tipo de organización visto antes no es adecuado. Esto se debe a que las variables cuantitativas por lo regular presentan muchos valores distintos, con lo cual la finalidad de condensar la información no se cumple.
La idea ahora consiste en establecer intervalos que cubran todos los datos que se tienen a disposición sobre la variable en estudio. De esta manera, se construye una tabla en la que se cuenta el número de observaciones contenidas en cada intervalo previamente especificado. Estos intervalos se llaman clases o intervalos de clases y el número de datos en cada intervalo se denomina frecuencia. Esta forma de agrupar los datos tendrá esta apariencia:

Intervalos de clase	**frecuencia**
LI_1 - LS_1	f_1
LI_2 - LS_2	f_2
...	...
LI_i - LS_i	f_i
...	...
LI_k - LS_k	f_k
	Total de observaciones

De este modo, se puede definir una **distribución de frecuencias** como una ordenación tabular de los datos en intervalos de clase con sus respectivas frecuencias.

Nota:
Cuando los datos se presentan en distribuciones de frecuencias, se habla de *datos agrupados*, mientras que cuando se presentan individualmente, se habla de *datos no agrupados*.

Pasos para la construcción de una distribución de frecuencias

1. Determinar el valor máximo y el valor mínimo de los datos.
2. Calcular el rango (o recorrido) de la variable el cual viene dado por la diferencia entre el valor máximo y el valor mínimo. El rango se denota por ***R***.
3. Determinar el *número de clases* (***K***) y las *amplitudes de clase* (***C_i***):
   A la anchura de un intervalo de clase se le conoce como *amplitud de clase*, es decir, la amplitud de clase de un intervalo viene dada por la diferencia entre el límite superior y el límite inferior de dicho intervalo. Podemos determinar la amplitud o el número de clases tomando en cuenta lo siguiente:

   a. Si se conoce el número de clases: $C_i = \frac{R}{K}$

   b. Si se conoce la amplitud de las clases: $K = \frac{R}{C_i}$

c. Regla de Sturges:
*K = 1 + 3,3 * Log n*

Nota:

i. La fórmula de Sturges sólo proporciona una orientación sobre cuál debe ser el número de clases. También se puede usar la regla de la raíz cuadrada: $K = \sqrt{n}$.

ii. Pueden existir clases abiertas, es decir, clases que sólo tienen un límite superior o solamente un límite inferior. Si ese es el caso, a esta clase abierta no se le podrá determinar la amplitud.

iii. En la práctica no se conoce de antemano el número de clases y la amplitud de estas. Sin embargo existen dos recomendaciones importantes al construir una distribución de frecuencias:

- Que el número de clases no sea inferior a 5 ni mayor que 15.
- De ser posible es deseable que todas las clases tengan la misma amplitud.

4. Proceder a construir los intervalos de clase.
En este punto ya se debe conocer el número de intervalos de clase a construir y las amplitudes de clase de cada uno de ellos, las cuales pueden ser iguales o no. Para la construcción de las clases se deben seguir los siguientes pasos:
   a. Establecer el límite inferior del primer intervalo de clase. Esto se puede realizar arbitrariamente de acuerdo a las siguientes alternativas:
      - Utilizando el valor mínimo de los datos
      - Utilizando otro valor menor al mínimo, pero no muy alejado.
   b. Fijado el primer límite inferior se le suma a este la amplitud de la primera clase, C_1, y se obtiene el límite superior de esta primera clase, el cual se constituye a la vez como el límite inferior de la segunda clase, a este se le suma la amplitud C_2 y se obtiene el límite superior de la segunda clase. Y de la misma manera se construyen los K intervalos. Naturalmente el último intervalo de clase debe incluir el valor máximo de los datos.
   c. Para calcular la frecuencia de cada intervalo, se debe asumir lo siguiente: En términos matemáticos los intervalos de clase van a ser intervalos cerrados por su límite inferior y abiertos por su límite superior. Es decir, el intervalo de la i-ésima clase será $[LI_i - LS_i)$, con $i = 1, \dots, K$.

5. Determinar el número de datos contenidos en cada clase. Es decir, determinar las *frecuencias absolutas* de clase (f_i). Evidentemente se debe cumplir que $\sum_{i=1}^{K} f_i = n$, siendo n el número total de datos

6. Determinar el resto de las frecuencias.
   a. Frecuencia relativa de una clase:
   Se va a denotar por fr_i y se obtiene de la siguiente manera:

$$fr_i = \frac{f_i}{n}$$

Siempre se cumple que $\sum_{i=1}^{K} fr_i = 1$. La frecuencia relativa de una clase representa la *proporción de datos* contenidos en ese intervalo de clase.

b. Frecuencia acumulada de una clase:
Se denota por F_i. Se obtiene sumando las frecuencias absolutas de todas las clases anteriores a ella más la frecuencia absoluta de la *i*-ésima clase considerada. Por tanto la frecuencia acumulada de la última clase es $F_k = n$.
La frecuencia acumulada, F_i, representa el *número de observaciones que son menores* que el límite superior de la *i*-ésima clase.

c. Frecuencia relativa acumulada de una clase:
Se denota por Fr_i y se obtiene de la siguiente manera:

$$Fr_i = \frac{F_i}{n}$$

También, $Fr_i = fr_1 + fr_2 + fr_3 + \ldots + fr_i$
La frecuencia relativa acumulada, Fr_i, representa la *proporción de todas las observaciones que son menores* que el límite superior de la i-ésima clase.

d. Marca de clase o punto medio de clase:
La *marca de clase* o *punto medio de clase,* denotado por m_i se define como el punto central de la clase particular:
$m_i = \frac{LI_i + LS_i}{2}$, en donde LI_i es el límite inferior de la *i*-ésima clase y LS_i es el límite superior de esa clase.

Ejercicio:
Con base en los datos recogidos en clase, construir una distribución de frecuencias para la variable peso.

Nota 3:
Existen algunas situaciones en que uno o más intervalos de clase en una distribución de frecuencias no tienen límite inferior o superior. Estos se conocen como ***Clases Abiertas***.
Por ejemplo, la siguiente distribución de frecuencias tiene dos clases abiertas:

Clases	fi
Menos de 5	73
5 - 10	58
10 - 15	35
…	…
50 y más	39

Observe que a las clases abiertas no se les puede determinar la amplitud y tampoco la marca de clase.

Ejercicio:
Completar la siguiente distribución de frecuencias:

Clases	m_i	f_i	fr_i	F_i	Fr_i
- - -	15	--	---	--	0,16
[20 -30)	--	--	0,08	--	---
[30 - 40)	--	6	---	12	---
[40 - 50)	45	8	---	20	---
- - -	65	--	---	--	---
Totales		--	---		

Distribución de frecuencias cuyas clases son valores individuales de la variable en estudio

En muchas ocasiones se presentan colecciones de datos en las cuales el número de valores diferentes que toma la variable de interés es pequeño y por consiguiente, no es apropiado agrupar estos datos en una distribución de frecuencias cuyas clases sean intervalos. Generalmente, en estos casos, los datos son de tipo discreto.
En tal situación, se toman como clases los diferentes valores de la variable y las frecuencias se calculan de la forma habitual.

Ejemplo:
Con base en los datos recogidos en clase, construir una distribución de frecuencias para la variable *Número de veces al mes que va al cine*.

Nota:
a. Nótese que en este tipo de distribución de frecuencias no existen límites de clase, amplitudes y las marcas de clase m_i coinciden con las clases.
b. Obsérvese también que en las distribuciones de frecuencias cuyas clases son valores individuales, se puede reconstruir fácilmente la colección de datos originales. Recuerde que esto no es posible cuando las clases son intervalos.

Ventajas y desventajas de agrupar los datos en distribuciones de frecuencias

- Facilita la presentación y resumen de los datos, lo que permite analizar sus aspectos más resaltantes.
- La desventaja principal es que se pierde la individualidad de los datos. Se sabe que en determinado intervalo está contenido cierta cantidad de datos pero no se conoce exactamente qué valores toman.

En conclusión, al agrupar datos se gana en simplicidad y accesibilidad, pero se pierde el nivel de detalle en el caso de distribuciones de frecuencias con intervalos.

Distribuciones de frecuencias para datos Cualitativos

Las distribuciones de frecuencias también se pueden utilizar para datos cualitativos. Éstas son más fáciles ya que las clases se ponen de manifiesto con más facilidad, de tal manera que los cálculos son mínimos.
Por ejemplo, en la siguiente tabla se presentan las ventas de gaseosas, ordenadas en una tabla de frecuencia:

Sabor	**fi**	**fri x 100**
Cola	600	60%
Limón	200	20%
Naranja	100	10%
Uva	50	5%
Fresa	40	4%
Otros	10	1%
	1000	100%

Observe que no se calculan las frecuencias acumuladas. Esto se debe a que en este caso no tiene sentido dado que los valores de la variable se pueden ordenar de forma arbitraria.

Ejercicio:
Calcule la frecuencia acumulada a la distribución de frecuencias anterior e intente interpretarla de manera similar a como lo hizo con la variable peso. Debe notar que tal interpretación carece de sentido en este caso particular.