Distribuciones de Frecuencias

Tema 3

Profesor Tevni Grajales G.

Cuando los datos que deben analizarse son numerosos, como puede ser la edad o el salario mensual de una población de tres mil personas, es conveniente agrupar los datos de manera ordenada por clases o categorías que muestran, para cada una de ellas, el número de elementos que contiene o frecuencia. A esto es lo que se denomina, distribución de frecuencia.

Como ejemplo tenemos a continuación los datos relacionados con el CI de 110 estudiantes seleccionados al azar en una institución.

Tabla No.1

CI de 110 estudiantes seleccionados al azar

154	131	122	100	113	119	121	128	112	93
133	119	115	117	110	104	125	85	120	135
116	103	103	121	109	147	103	113	107	98
128	93	90	105	118	134	89	143	108	142
85	108	108	136	115	117	110	80	111	127
100	100	114	123	126	119	122	102	100	106
110	113	112	82	114	112	113	142	145	123
97	135	108	139	133	107	115	83	109	116
118	104	127	94	115	101	125	129	131	110
150	130	87	89	108	137	124	96	111	101
105	111	127	108	106	91	123	132	97	110

Como podemos notar, este conjunto de datos están presentados de tal manera que resulta muy difícil poder visualizar su comportamiento, surge pues la necesidad de organizarlos o sistematizarlos para una mejor compresión. Un primer paso podría ser ordenar los valores de mayor a menor e identificar cuántas veces se repite cada valor. Eso como cuando realizamos votaciones en el aula de clase y ponemos una rayita al lado del nombre de cada candidato en la medida como consigue un voto.

En la Tabla No. 2 se presentan estos mismos resultados organizados de mayor a menor. Cada valor esta ubicado bajo la columna X y la cantidad de casos por cada valor está representado bajo la columna f con una diagonal.

Tabla No.2

Distribución de frecuencias de CI de 110 estudiantes

X	f	X	f	X	f	X	f	X	f	X	f
154	/	140		126	/	112	///	98	/	84	
153		139	/	125	//	111	///	97	//	83	/
152		138		124	/	110	/////	996	/	82	/

151		137	/	123	///	109	//	95		81	
150	/	136	/	122	//	108	//////	94	/	80	/
149		135	//	121	//	107	//	93	//		
148		134	/	120	/	106	//	92			
147	/	133	//	119	///	105	//	91	/		
146		132	/	118	//	104	//	90	/		
145	/	131	//	117	//	103	///	89	//		
144		130	/	116	//	102	/	88			
143	/	129	/	115	///	101	//	87	/		
142	//	128	//	114	//	100	////	86			
141		127	///	113	////	99		85	//		

Tenemos pues una distribución "simple" de frecuencias de resultados, se puede identificar que los datos están dispersos, algunos valores no tienen frecuencias de ocurrencia. Aunque ahora es posible mejor entender los datos, estos todavía parecen muy numerosos para observarlos. Por eso algunos investigadores acostumbran reagruparlos en lo que se denomina distribución de frecuencias de "datos agrupados". La Tabla No. 3 nos presenta los mismos resultados presentados en distribución de frecuencia en datos agrupados. Como usted puede notar, se establecen categorías o clases que agrupan un segmento de valores. Estos segmentos son denominados intervalos de clase, el tamaño del intervalo de clase está determinado por la extensión de los valores, es decir si la diferencia entre más grande sea la diferencia entre el mayor valor y el menor, más grande debe llegará a ser el intervalo. Pero esta no es una regla, pues hay que tomar en cuenta que se debe conservar la capacidad de discriminación proporcionada por la medida original, tampoco deben ser tan pequeños los intervalos al grado de que se pierda el propósito que tenemos de facilitar el manejo de los datos. En muchos casos se sugiere que no se establezcan hasta 15 intervalos de clase. Para calcular el tamaño del intervalo, reste el valor más bajo del más alto y súmele una unidad. En el caso de nuestro ejemplo (154 – 80) + 1 = 75 esto nos permite conocer el total de datos potenciales. Una vez realizada esta operación, divida este resultado entre 15 suponiendo que vamos a tener 15 categorías o clases, de esa manera tendremos el tamaño que corresponde a cada intervalo de clase. 75/15 = 5. Lo que significa que la amplitud del intervalo de clase será de cinco (i = 5).

Tabla No 3

CI de los estudiantes presentado en distribución de frecuencias agrupadas

Intervalo de clase	f	Intervalo de clase	f	Intervalo de clase	f
150 - 154	2	125 - 129	9	100 – 104	12
145 - 149	2	120 – 124	9	95 – 99	4
140 - 144	3	115 – 119	13	90 – 94	5
135 - 139	5	110 – 114	17	85 – 89	5
130 - 134	7	105 – 109	14	80 – 84	3
					N=110

Su toma el valor más bajo como límite inferior del primer intervalo y se le suma el valor i y se le resta l. En este caso el valor más bajo es 80 y se le suma i = 5 dando como resultado 85 y al restarle uno tenemos el límite superior del intervalo que es 84. Así es como resulta el intervalo 80 – 84 que aparece hacia el extremo inferior de la tercera (última) columna. Se sigue el mismo procedimiento para determinar los restantes intervalos de clase. Una vez concluido este proceso se identifica el número de casos o frecuencia de casos que están incluidos en el intervalo correspondiente. Para ello puede hacer uso de la Tabla No. 2. Un detalle que se debe mencionar es que el "verdadero" valor de un número es igual a su valor aparente más y menos la mitad de la unidad de medida. De manera que si queremos indicar los límites verdaderos de los intervalos de clase se determinan con el valor aparente menos la mitad de la unidad de medida del valor más bajo en el intervalo para determinar el límite real inferior del intervalo y el valor aparente más la mitad de la unidad de medida del valor más alto para el límite real superior. En el ejemplo anterior, para el primer intervalo tendremos 79.5 límite real inferior y 84.5 límite real superior.

Las distribuciones de frecuencia en variables discretas dan lugar a otras informaciones valiosas. A fin de ilustrarlo a continuación algunos datos recogidos entre un grupo de 25 alumnos a quienes se les preguntó cuántos hermanos tienen. Una vez determinada la cantidad de hermanos por cada alumno, se agrupó en una distribución de frecuencia simple según aparece en la Tabla No. 4 En la primera columna se presenta el total de hermanos organizados en clase, luego la columna dos presenta el recuento y la columna tres la frecuencia absoluta de estudiantes que tienen un total de hermanos que coincide con la respectiva clase. Luego se calculó la frecuencia relativa (columna cuatro), dividiendo la frecuencia absoluta de cada clase entre el total de observaciones (casos), lo cual permite conocer qué proporción de los alumnos tienen un número determinado de hermanos.

Para ciertos propósitos resulta útil acumular las frecuencias absolutas –o relativas -. Esto puede hacerse sumándolas hacia abajo, para obtener una distribución acumulada *"menos de"* o hacia arriba, lo que produce una distribución acumulada *"más de"* según aparecen en las columnas cinco y siete para las absolutas, seis y ocho para las relativas.

Tabla 4

Datos agrupados, frecuencias absolutas y relativas a más y menos de

Número de hermanos	Recuento	Frecuencia absoluta (f)	Frecuencia relativa (fr)	Acumulada absoluta *"menos de"*	Relativa *"menos de"*	Acumulada absoluta *"más de"*	Relativa *"más de"*
1	//	2	0,08	2	0,08	25	1,00
2	///	3	0,12	5	0,20	23	0,92
3	////	4	0,16	9	0,36	20	0,80
4	////////	8	0,32	17	0,68	16	0,64
5	///	3	0,12	20	0,80	8	0,32
6	/	1	0,04	21	0,84	5	0,20
7	//	2	0,08	23	0,92	4	0,16
8	//	2	0,08	25	1,00	2	0,08

Si se suman las frecuencias absolutas de las clases 1ª, 2ª, 3ª, 4ª, se obtiene 2 +3 + 4 + 8 = 17, valor que representa la frecuencia acumulada "menos de" de la cuarta clase. Si se realiza la misma operación con las frecuencias relativas se obtiene 0,68. Lo cual quiere decir que 17 estudiantes tienen 4 hermanos o menos (recuérdese que se sumaron las categorías 1,2,3,4); aunque también sería correcto decir que 17 estudiantes tienen *menos de* 5 hermanos. La interpretación del 0,68 es similar; 68% de los estudiantes tienen 4 hermanos o menos.

Si la suma se realiza de abajo hacia arriba, partiendo de la última clase se obtiene una frecuencia acumulada "más de". Los números 5 y 0,20 de la sexta clase son ejemplos de ellas. Estos valores se obtuvieron de la forma siguiente. El número 5 se obtuvo sumando las frecuencias absolutas de las clases 8ª, 7ª, 6ª, o sea 2 + 2+ 1 = 5 y el número 0,20 sumando 0,08 + 0,08 + 0,04 = 0,20. Su interpretación se deriva de su forma de cálculo. El número 5 nos indica que 5 estudiantes tienen 6 hermanos o más, aunque sería correcto decir que 5 estudiantes tienen más de 5 hermanos. El número 0,20 se interpreta en forma similar: 20% de los estudiantes tienen 6 hermanos o más.

Este ejercicio nos lleva al concepto estadístico denominado centil (en este caso se divide la distribución en centésimos), se trata de un punto en una escala de puntajes debajo del cual caen cierto porcentaje de casos. Por ejemplo, Pedro ha sido calificado en el centil 45 en una prueba de matemáticas. Esto significa que el 45% de los puntajes obtenidos por los alumnos están por debajo de él mientras que el restante 55% se ubica por encima de él. Si en el mismo caso Amelia obtuvo el centil 68, podemos decir que ella se ubica sobre el 68% de las personas que contestaron la prueba y que el 32% logró una calificación superior a la que ella obtuvo. También se utilizan términos como "deciles" los cuales nueve puntos dividen la distribución en diez partes y cuartiles son tres puntos que dividen la distribución en cuatro porciones. De manera que el primer cuartil corresponde al centil 25 el segundo cuartil al centil 50 (también conocida como mediana), el tercer cuartil corresponde al centil 75. Se utiliza también el término distancia intercuartil para referirse a los valores que se ubican entre el primero y el tercer cuartil.

Los centiles son utilizados con mucha frecuencia para interpretar los resultados de pruebas estandarizadas. Entre sus limitaciones se puede mencionar que no pueden ser utilizados para hacer sumas y obtener promedios entre dos o más distribuciones a menos que sus medidas de tendencia central y variabilidad sean iguales. La diferencia real que puede existir entre los puntos obtenidos por una persona y otra puede ser mucho más grande de lo que aparenta en una escala de percentiles. De manera que la diferencia en puntos que puedan tener las personas en centil 40 y 50 no necesariamente es la misma diferencia en puntos entre las personas que se ubican en el centil 60 y 70 de la misma distribución. Lo que lleva a que los centiles tienden a minimizar las grandes diferencias entre los puntajes especialmente en los extremos de la distribución y tiende al mismo tiempo a magnificar las diferencias pequeñas entres los puntajes obtenidos al centro de la distribución.

Para calcular los centiles hay diferentes procedimientos dependiendo de las características de la distribución. Cuando no hay puntajes y se trata de menos de 30 casos, se ordenan los casos por rango (valor 1 para el más bajo) luego se calcula el centil restando al valor del rango de cada caso 0,5 y dividiéndo el resultado entre el número de casos. Cuando los casos son 30 o más se divide el valor del rango entre el total de casos. Cuando hay más de una persona que tiene el mismo puntaje se utiliza la frecuencia acumulada a *menos de* y se suma mitad de frecuencias en hay en la clase que se observa y se divide entre el total de casos.