TABLAS DE DISTRIBUCIÓN DE FRECUENCIA

En el momento que se manejan los datos de una muestra o de una población y el conjunto tiene 20 o más observaciones, la mejor manera para examinar esos datos es presentarlos en forma resumida, elaborando las tablas o cuadros apropiados, que proporcionan una base para su representación gráfica, facilitan el cálculo de las diferentes medidas y el análisis de las principales características de la información recogida. Para elaborar los cuadros, se debe, antes que todo, identificar las variables que intervienen. Cuando se trabaja con dos variables se denomina distribución bidimensional (ver tabla).

Sucursal	Valor (miles $)
Centro	225
Chapinero	105
Chicó	150
Kennedy	221
Lago	312
Restrepo	57
Total	1070

Tabla 1. Distribución bidimensional

En el caso de estudiar tres o más características simultáneamente, se trata de una distribución **pluridimensional o multidimensional**; por ejemplo al clasificar las ventas de una empresa por sucursales, trimestres y valores :

Sucursal	Ventas (millones$) trimestres 1, 2 ,3 y 4				Total
Centro	18,1	22	17,5	24	
Chapinero	10	10,4	9,8	13	43,2
Chicó	13,7	12	12,8	16	54,5
kennedy	20	17,5	18,3	22	77,8
Lago	27	24,5	25,9	29	106,4
Restrepo	5,35	5,68	6,28	7,2	24,51
TOTAL	94,15	92,08	90,58	111,2	388,01

Tabla 2. Distribución pluridimensional

Distribuciones de frecuencia acumulada

Para realizar una visión inicial de la información y a manera de ordenamiento estratificado se hace tabla de frecuencias, así los datos se clasifican y ordenan de

acuerdo con ciertas características cualitativas y cuantitativas, mostrando el número de veces que se repite la variable o la característica cualitativa.

Vamos a aprender cómo se realiza esta tabla de frecuencias, pero antes se hace necesario verificar simbología usada en este tipo de tablas.

- **N** = tamaño poblacional
- **n** = tamaño de muestra
- $\mathbf{X_1}$ = característica cuantitativa, observada en cada unidad investigada
- $\mathbf{f_i}$ = frecuencia absoluta. Número de veces que se repite cada valor de la variable.
- **% o $\mathbf{h_r}$** = f_i/n frecuencia porcentual o relativa. (Se usa porcentual si se quiere trabajar con una base de 100, o relativa si se quiere trabajar con una base de 1) Se halla dividiendo cada frecuencia absoluta entre la población. (Para indicar el porcentaje al anterior resultado se multiplica por 100%)

La forma de hacer una tabla de frecuencias: supongamos que hay 300 cajas de cartón en un salón y cada una de ellas contiene platos de vidrio. Se desea examinar las cajas con el fin de saber el número de platos que han sufrido desperfectos en el transporte, desde la fábrica hasta el salón. Por motivos de tiempo, espacio y personal disponible, se toma la decisión de revisar un 10% de las cajas; por tal razón se tendrá 30 cajas de un total de 300.

- ***N*** = 300 (tamaño de la población objetivo)
- ***n*** = 30 (tamaño de la muestra)

Cada caja seleccionada, en forma aleatoria, se simboliza por x_i (minúscula en la muestra, mayúscula en la población), donde el subíndice **i** toma valores desde uno hasta **n**, siendo x_1 la primera caja seleccionada, x_2 la segunda, y así sucesivamente. Cada x_i tendrá como valor el correspondiente a la característica examinada; en este ejercicio le corresponderá el número de platos de porcelana desperfectos.

$x_1 = 2$	$x_2 = 1$	$x_3 = 1$	$x_4 = 0$	$x_5 = 3$
$x_6 = 3$	$x_7 = 2$	$x_8 = 1$	$x_9 = 2$	$x_{10} = 4$
$x_{11} = 0$	$x_{12} = 2$	$x_{13} = 3$	$x_{14} = 1$	$x_{15} = 2$
$x_{16} = 2$	$x_{17} = 3$	$x_{18} = 0$	$x_{19} = 3$	$x_{20} = 2$
$x_{21} = 2$	$x_{22} = 2$	$x_{23} = 1$	$x_{24} = 4$	$x_{25} = 3$
$x_{26} = 2$	$x_{27} = 3$	$x_{28} = 2$	$x_{29} = 1$	$x_{30} = 2$

Tabla 3. Número de platos despefectos

Número de platos	Tabulación	# de cajas (fi)	hi

Desperfectos	1a. Forma	2a. Forma		
0	///	ëê	3	0,10
1	𝍸 //	ë	6	0,20
2	𝍸 𝍸 𝍸		12	0,40
3	𝍸 ///	ëê	7	0,23
4	//	ë	2	0,07
TOTAL			**30**	**1,00**

Tabla 4. Datos sin agrupar.

En las columnas de tabulación se presentan dos formas de realizar el conteo manual sobre el número de veces que se presenta cada valor que toma la variable. Cada raya corresponde a una observación, evitando hacer acumulados de rayas. De esta manera /////, que luego al ser contados pueden dar lugar a equivocaciones, de ahí que sea preferible formar grupos de cuatro rayas (𝍸 ó ⌊) con lo cual disminuye la posibilidad de error que se puede presentar al hacer el recuento en grupos grandes.

Para la presentación de un informe se había anotado que todo cuadro requiere enumeración si hay varios; además del título completo que indique su contenido.

Prescindiéndose de las columnas que utilizamos para la tabulación de la tabla 4, reemplazadas por la frecuencia absoluta, con la posibilidad de agregar otra columna, correspondiente a la frecuencia relativa, la que nos indicará la distribución porcentual. En el mismo cuadro, por ejemplo: se tendrá que el 10% de las cajas no tienen platos defectuosos, porcentaje que se obtiene de dividir la frecuencia absoluta (tres) por el tamaño de la muestra (treinta) y luego multiplicar por 100 así:

$$\frac{3}{30} X100 = 10\%$$

Es importante destacar, en el caso de variables cualitativas, pueden ser analizadas, en parte, mediante el cálculo de porcentajes, y, al igual que las variables cuantitativas, se pueden representar gráficamente.

A medida que se incrementa el número de observaciones, se hace necesario condensar los datos en tablas apropiadas de resumen. Para ello, se acomodan los datos en grupos (**intervalos**) de clases (es decir, **categorías**) dividiendo en forma conveniente las observaciones. A este arreglo de datos en forma tabular se le denomina distribución de frecuencia.

Una distribución de frecuencia es una tabla-resumen en la que se disponen los datos divididos en grupos ordenados numéricamente, y que se denominan clases o

categorías. Cuando se "agrupan" o se les condensa en tablas de distribución de frecuencia, es más manejable y significativo el proceso de análisis e interpretación de datos. En esa forma resumida es muy sencillo aproximar las principales características de los datos y de esta manera se compensa el hecho de que, al agrupar los datos se pierde alguna información inicial referente a las observaciones individuales. Para mejorar el análisis es deseable indicar en la tabla la frecuencia porcentual principalmente cuando se compara un conjunto de datos con otro y, en especial, si es distinto el número de observaciones de cada conjunto. Al construir la tabla de distribución de frecuencia se debe prestar atención en:

1. Seleccionar el número adecuado de clases para la tabla.
2. Obtener un intervalo de "anchura" apropiado.
3. Establecer los límites de cada clase para evitar traslapes.

Ejemplo variable continua

Consideramos la población de las cajas (N = 300) y seleccionemos aleatoriamente una muestra de 30 cajas (n = 30), o sea el 10%, a fin de investigar el peso en Kg. de cada caja, se da en números enteros con el fin de simplificar el trabajo, sin olvidar que la medida (peso) utilizada admite valores fraccionarios (kilogramos y gramos), por tal motivo se le clasifica como variable continua.

$x_1 = 48$	$x_2 = 56$	$x_3 = 60$	$x_4 = 67$	$x_5 = 47$
$x_6 = 70$	$x_7 = 70$	$x_8 = 63$	$x_9 = 72$	$x_{10} = 76$
$x_{11} = 74$	$x_{12} = 67$	$x_{13} = 92$	$x_{14} = 70$	$x_{15} = 69$
$x_{16} = 61$	$x_{17} = 71$	$x_{18} = 79$	$x_{19} = 85$	$x_{20} = 68$
$x_{21} = 82$	$x_{22} = 55$	$x_{23} = 65$	$x_{24} = 88$	$x_{25} = 52$
$x_{26} = 58$	$x_{27} = 76$	$x_{28} = 57$	$x_{29} = 72$	$x_{30} = 67$

Tabla 5. Datos sin agrupar

En la elaboración de la tabla o cuadro de frecuencias, se realizan los siguientes pasos:

1) El primer paso a seguir es determinar el valor máximo y mínimo que toma la variable x_i:

$$X_{\min} = 41 \qquad X_{\max} = 92$$

2) La diferencia que hay entre el valor máximo y el mínimo se denomina rango o recorrido. En este ejemplo será:

$$X_{\max} - X_{\min} = 47$$

3) Se hace necesario determinar el número de intervalos (m) que se utilizará para agrupar los datos:

m = número de intervalos o de clases

Una de las formas de obtener m es aplicando la regla de Sturges, con la cual se obtiene una aproximación aceptable sobre el número de intervalos necesarios. Aplicando dicha fórmula al ejercicio se tendrá:

$$m = 1 + 3,3 \log 30$$
$$m = 1 + 3,3(1,4771)$$
$$m = 1 + 4,8745 = 5,87$$
$$m \approx 6$$

El número de intervalos de acuerdo a la regla de Sturges, estará entre 5 y 6. Utilizaremos en nuestro ejercicio seis intervalos (m = 6).

En la práctica **m** se determina atendiendo varios factores, tales como: finalidad del estudio, grado de variabilidad de los datos, necesidad de efectuar comparaciones. En todo caso, se recomienda que el valor de m, hasta donde sea posible, no sea menor de 5, ni mayor de 16. Si no existen suficientes clases, o si hay demasiadas, la información que se puede obtener es reducida.

4) Una vez determinado el número de intervalos, se debe decidir sobre el valor de la amplitud para cada intervalo:

C = amplitud del intervalo

Al determinar el valor de C, no es necesario que sea igual para todos los intervalos, tal como acontece en numerosos casos prácticos. Sin embargo, con fines de simplificaciones y de funcionalidad, se puede considerar el valor de C constante para todos los intervalos. Dicho valor constante se obtiene aplicando la fórmula siguiente:

$$C = \frac{rango}{m} = \frac{X_{\max} - X_{\min}}{\#\ de\ \text{int}ervalos}$$

En nuestro ejercicio se tendrá:

$$C = \frac{X_{\max} - X_{\min}}{\#\ de\ \text{int}ervalos} = \frac{92 - 47}{6} = 7.5$$

Para facilitar los cálculos se aproximaría C a 8; por lo tanto se altera el valor del rango. Si recordamos que m ya fue hallado y no se desea cambiar se tendrá:

Anteriormente: $7.5 = \frac{45}{6}$

Ahora: $8 = \frac{Rango}{6} \therefore 8 = \frac{48}{6}$

El rango se incrementa en tres unidades, de 45 pasó a 48. El incremento debe ser distribuido proporcionalmente, sumando unas unidades al límite superior y restándole otras al límite inferior. Las situaciones que se pueden presentar al hacer la repartición del incremento se exponen a continuación.

Cualquiera de las situaciones siguientes en la determinación de los límites del nuevo rango son válidas, siendo preferible distribuir dicho incremento en forma proporcional.

X_{max}	X_{min}	Recorrido	
92	**47**	**45**	**Original**
93	**45**	**48**	
94	**46**	**48**	**Nuevo Rango**
95	**47**	**48**	
92	**44**	**48**	

Tabla 5. Datos sin agrupar

Esta es la razón por la cual se tomará como:

X_{max} a 94 y X_{min} a 46

5) La columna correspondiente a la variable continua se simbolizará:

$\boldsymbol{X'_{i-1} - X'_{i}}$: (ambas minúsculas para la muestra y en la población deberán ser mayúsculas).
X'_{i-1}: Límite inferior del intervalo
X'_{i} : Límite superior del intervalo.

6) La tabla sobre frecuencias se basa en la información correspondiente al peso de cada una de las 30 cajas examinadas.
Para la elaboración de los intervalos, se inicia con la determinación del valor x_{Min} en el nuevo rango, siendo en nuestro caso 46, el cual se toma como límite inferior (x_0) del primer intervalo, luego se procede a agregarle el valor de la amplitud para así obtener el límite superior (x_1), que será a su vez el límite

inferior del segundo intervalo, al cual se le agrega nuevamente el valor de C para obtener el límite superior del segundo intervalo y así sucesivamente hasta conformar la columna de la variable continua.

Peso (kg) Intervalos $Y'_{i-1} - Y'_i$	Registro de la frecuencia	Frecuencia absoluta n_i	Frecuencia relativa h_i	Frecuencia acumulada n_i	Frecuencia relativa acumulada h_i	Marca de clase y_i
46,1- 54	||	3	0,10	3	0,10|	50
54,1-62	|	6	0,20	9	0,30	58
62,1-70	|	10	0,33	19	0,63	66
70,1-78	|	6	0,20	25	0,83	74
78,1-86	||	3	0,10	28	0,93	82
86,1-94	|	2	0,07	30	1,00	90
$\sum$		30	1,00	--	--	--
Peso (kg) Intervalos $X'_{i-1} - X'_i$	**Registro de la frecuencia**	**Frecuencia absoluta** f_i	**Frecuencia relativa** f_i/n	**Frecuencia acumulada** F_i	**Frecuencia relativa acumulada** F_i/n	**Marca de clase** X'_i

Tabla 6. Tabla elaboración de frecuencias

Se observará también que a cada uno de los límites inferiores de los intervalos se les agregó 0,1, con el fin de facilitar la clasificación de cada observación, así por ejemplo X6 = 70 estaría considerada en el intervalo 62,1 -- 70 y no en el intervalo de 70,1 -- 78, procedimiento que evita la dificultad al no saber dónde clasificar dicho valor al tener intervalos, tales como (62 -- 70) y (70 -- 78). Debe quedar bien claro que la amplitud del intervalo sigue siendo 8 y que el 0,1 es usado únicamente como ayuda para la clasificación.

Otras formas de clasificar la información de la tabla anterior pueden ser la siguiente:

Peso (kg) Intervalos $Y'_{i-1} - Y'_i$	Peso (kg) Intervalos $Y'_{i-1} - Y'_i$
46- 53,9	46- 52
54-61,9	53-59
62-69,9	60-66
70-77,9	67-73
78-85,9	74-80
86-93,9	81-87
	88-94
Peso (kg) Intervalos	**Peso (kg) Intervalos**

$X'_{i-1} - X'_i$	$X'_{i-1} - X'_i$

Tabla 7. Otras formas de clasificación

En la tabla anterior el valor de X = 70 quedará incluido en el intervalo 70 - 77,9. Aumentó el número de intervalos a 7 y el tamaño del intervalo pasó a ser 7, porque en este caso, C = (52-46) + 1=7. En la tabla de frecuencia la columna simbolizada por se denomina marca de clase, la cual sirve para facilitar el cálculo de algunas medidas de posición y de dispersión, la marca de clase o punto medio se puede obtener de tres formas diferentes:

1. Como promedio de los límites de cada intervalo, (tomando los datos de las tablas):

$$X_1 = \frac{X'_0 + X'_1}{2} = \frac{46+54}{2} = 50$$
$$X_2 = \frac{X'_1 + X'_2}{2} = \frac{54+62}{2} = 58$$
$$X_3 = \frac{X'_2 + X'_3}{2} = \frac{62+70}{2} = 66$$
$$\vdots \quad \vdots \quad \vdots \quad \vdots$$
$$X_6 = \frac{X'_5 + X'_6}{2} = \frac{86+94}{2} = 90$$

2. Si la amplitud (C) del intervalo es constante, basta con determinar la primera marca de clase, de acuerdo con el método anterior, luego se le va sumando el valor de la amplitud, tal como se presenta a continuación:

$$X_1 = \frac{X'_0 + X'_1}{2} = \frac{46+54}{2} = 50$$

$$X_2 = X_1 + C = 50 + 8 = 58$$

$$X_3 = X_2 + C = 58 + 8 = 66$$

$$X_4 = X_3 + C = 66 + 8 = 74 \text{ y así sucesivamente}$$

3. Otro método para hallar las marcas de clase (X_ii) consiste en dividir la amplitud de cada intervalo por dos; luego, este resultado se le suma al límite inferior del respectivo intervalo:

$$X_1 = X_0 + \frac{C}{2} = 46 + \frac{8}{2} = 50$$
$$X_2 = X_1 + \frac{C}{2} = 54 + \frac{8}{2} = 58$$

En una variable, ya sea discreta o continua, cuando las frecuencias absolutas o relativas equidistantes a un valor central son iguales, se dice que la distribución es **simétrica**, como se puede observar en la siguiente tabla:

X'_i	f_i	h_i
3	2	0,10
6	5	0,25
9	6	0,30
12	5	0,25
15	2	0,10
S	20	1,00

Tabla 8. Variable discreta

$X'_{i-1} - X'_i$	f_i	h_i
Menores que 30	8	0,10
30,1-46	12	0,15
46,1-54	20	0,25
54,1-70	12	0,25
70,1-78	20	0,15
78,1- mas	8	0,10
S	80	1,00

Tabla 9. Variable continua

Ejemplo

Los siguientes datos corresponden a la distribución de frecuencias de los gastos en publicidad (en millones de $) de 50 empresas comerciales, durante el primer mes de 1998. Dichos gastos se agruparon en cuatro clases de amplitud constante, de la cual se sabe, se pide completar la tabla de frecuencias siguientes:

$X'_{i-1} - X'_i$	X_i	f_i	F_i	h_i	H_i
	3,5	4			
			20		
		25			
-8,75					
S		50			

Solución

Según los datos se obtienen

$$X_i = 3,5 \; X_4' = 8,75 \; f_1 = 4 \; F_2 = 20 \; f_3 = 25$$

Utilizando la fórmula:

$$X_1 = X'_0 + \frac{C}{2}$$

Reemplazando $X_1 = X_0' + \frac{C}{2} \rightarrow X_0' + 0,5C = 3,5$

Reemplazando $X_4 = X_0' + 4C \rightarrow X_0' + 4C = 8,75$

Si al límite inferior del primer intervalo le sumamos la mitad de **C**, se obtendrá la primera marca de clase. Ahora, si al mismo límite le sumamos cuatro veces el valor de **C**, se obtendrá el límite superior del cuarto intervalo.

Se tienen dos ecuaciones con dos incógnitas. Ahora obtendremos el valor de **C**, multiplicado a la primera ecuación por - 1 y luego se la restamos a la segunda ecuación:

$$\begin{array}{r} -X_0' - 0,5C = -3,5 \\ X_0' + \quad 4C = \; 8,75 \\ \hline 3,5C = 5,25 \end{array}$$

Despejando:

$$C = \frac{5,25}{3,50} = 1,5$$

$$X_0' + 0,5C = 3,5$$
$$X_0' + 0,5(1,5) \rightarrow X_0' = 2,75$$

Luego se obtiene la tabla de resultados solicitada así:

$X'_{i-1} - X'_i$	X_i	f_i	F_i	h_i	H_i
2,75-4,25	3,5	4	4	0,08	0,08
4,25-5,75	5,0	16	20	0,32	0,40

5,75-7,25	6,5	25	45	0,50	0,90
7,25-8,75	8,0	5	50	0,10	1,00
S		50		1,00	