Documento educativo

Statistics and health at work
Descriptive statistics (III): Measures in grouped data.

Estadística y salud en el trabajo
Estadística descriptiva (III): Medidas en datos agrupados.

Juan Luis Soto Espinosa[1].

(1) FES Zaragoza, UNAM
Edificio de Posgrado, planta baja, cubículo 17, FES Zaragoza, Campus II. Av. Batalla de 5 de mayo esq. Fuerte de Loreto Col Ejército de Oriente. C.P. 09230 Iztapalapa, CDMX

Correo electrónico de contacto:soej@unam.mx

Fecha de envío: 08/02/2021
Fecha de aprobación: 03/05/2021

Introducción

En el tema publicado en el número anterior, cuando se tiene una cantidad relativamente pequeña de datos (entre 50 y hasta 500 datos) es posible realizar un tratamiento estadístico para datos no agrupados. Cuando los estudios y análisis de datos se realizan en poblaciones muy grandes, incluso en las muestras, la cantidad de datos involucrados puede ser de cientos o miles de datos. Cuando se trabaja con grandes cantidades de datos, estos se agrupan en conjuntos y se siguen técnicas que permiten el cálculo de los estadísticos descriptivos (tendencias central y dispersión) que en conjunto se conocen como TÉCNICAS DE ANÁLISIS DE DATOS AGRUPADOS.

Para poder trabajar con altos volúmenes de datos, el primer paso es construir es una tabla de frecuencias, a la cual se le llamará **"tabla de frecuencias agrupadas",** estás se usan para resumir grandes cantidades de datos y facilitan el cálculo de las medidas de tendencia central y de dispersión.

A continuación (Tabla 1), se presenta la estructura general de una *tabla de frecuencias agrupadas* donde se muestran los elementos que la constituyen

En estadística, una CLASE se define como un conjunto de MODALIDADES (o valores) en que se divide un conjunto, de forma que la longitud de clase de cada uno sea exactamente la misma.

Tabla 1: Estructura de un atabla de frecuencia

Nombre de la variable (Intervalos de clase)	Frecuencia absoluta f_k	Frecuencia relativa fr_k	Frecuencia absoluta Acumulada fa_k	Frecuencia relativa acumulada fra_k	Marca de Clase MC_k
$[a_1, a_2)$					
$[a_2, a_3)$					
⋮					
$[a_k, a_{k+1}]$					

El uso de clases permite resumir de una manera más entendible un conjunto de datos que contiene una gran cantidad de modalidades y permite presentar información general descriptiva equivalente. Para entender mejor esta técnica de análisis es necesario familiarizarse con una serie de conceptos de suma importancia.

El primer aspecto que hay que notar, es que la tabla está dividida en intervalos, los cuales se presentan en la primera columna, para definir el intervalo se requieren de dos valores: un valor inferior (conocido como límite inferior de clase) y un valor superior (conocido como límite superior de clase). En la Tabla 1 se muestra la notación utilizada.

Dado el intervalo $[a_1, a_2)$, el paréntesis significa que el valor extremo a_2 queda excluido, mientras que el corchete indica que el valor extremo a_1 necesariamente se incluye en el intervalo. Por ejemplo, en el intervalo $[11, 16)$ el valor extremo 11 está incluido en el intervalo

y el valor extremo 16 queda excluido, para el siguiente intervalo $[16, 21)$ el valor extremo 16 está incluido y el 21 queda excluido. A este tipo de clases se les llama abierta por la izquierda y cerrada por la derecha.

En un conjunto de datos, que se ordena de menor a mayor, el valor más bajo se conoce como LÍMITE INFERIOR, mientras que el más alto se conoce como LÍMITE SUPERIOR, La distancia que existe entre estos dos valores se conoce como RANGO.

Matemáticamente, el RANGO se obtiene restando el valor menor del valor mayor, es decir:

Rango = Límite superior - Límite Inferior

Si tenemos el siguiente grupo de valores:

Tabla 2: Valores para ejemplo 1

Valor	Valor	Valor	Valor	Valor
101	83	48	40	49
123	34	60	84	83
75	82	126	118	20
12	107	106	28	58
116	118	99	75	61
25	122	87	54	112
28	30	20	9	70
83	115	108	86	121
50	45	41	89	42
92	107	76	26	130

Notemos que el LÍMITE INFERIOR es 10 (el valor más bajo) y el LÍMITE SUPERIOR es 130 (el valor más alto), por lo que el valor del Rango se obtiene realizando:

RANGO = LÍMITE SUPERIOR - LÍMITE INFERIOR

RANGO = 130 – 10

RANGO = 120

De donde se obtiene que la distancia que separa al valor más alto del más bajo es de 120 unidades.

Una vez que se conoce el rango, es necesario determinar el número de clases que se ha de considerar en la tabla de frecuencia. Una CLASE se define como un subconjunto de elementos (generalmente del mismo tamaño) en los que se dividen los datos ordenados provenientes de la población o muestra y que presentan características comunes.

El número de clases se identifica con la letra K; existen diversas formas para definir cuántas clases se deben considerar en la elaboración de una tabla de frecuencia; revisemos tres de ellas.

Primera: Considerar una tabla guía. Diversos autores han propuesto tablas para la elaboración de histogramas y selección de números de clases, por ejemplo, la propuesta por Roberto Behar y Pere Grima, la cual propone:

Tabla 3: Número de clases recomendadas según número de datos

Cantidad de datos	No. de clases
20 a 50	7
50 a 75	10
75 a 100	12
Más de 100	15

Otros autores sugieren 4 clases si tenemos entre 0 y 50 datos, 7 clases si tenemos entre 50 y 100 datos, 10 clases para más de 100 pero menos de 150 datos, 12 clases para más de 150 y menos de 200 datos y 14 clases para más de 200 datos.

Segunda: En ocasiones se recomienda determinar el número de clases a través de obtener la raíz cuadrada de la cantidad de datos. El resultado redondeado será el número de clases. La fórmula a resolver será:

$$\boldsymbol{K = \sqrt{N}}$$

Dónde:

K = Número de clases

N = Número de datos

Tercera: La opción matemáticamente más consistente es la conocida como **REGLA DE STURGES**, propuesta en el año de 1926 por el matemático Hebert Sturges. La solución de esta ecuación nos proporciona una regla práctica para obtener el número de clases:

$$\boldsymbol{K = 1 + 3.322 \log (N)}$$

Dónde

K = Número de clases

Log(N) = Logaritmo base 10 del número de datos

N = Número de datos

Una vez conocidos el rango y el número de clases a considerar, se debe determinar la amplitud de clase o ancho del intervalo. Se define como INTERVALO la distancia que existe entre los límites superior e inferior de una clase; se identifica con la con la letra h.

$$h = \frac{Rango}{K} = \frac{X_{Max} - X_{Min}}{k}$$

Dónde

h = Amplitud de intervalo

$\mathbf{X_{max}}$= Valor máximo del conjunto de datos

$\mathbf{X_{min}}$= Valor mínimo del conjunto de datos

K = Número de clases

En la medida de lo posible y si el estudio lo permite, el valor de la amplitud puede redondearse a un número entero para facilitar el cálculo de la longitud de cada intervalo, siempre y cuando el valor decimal sea mayor a 0.5 y cercano a la unidad siguiente.

Para tener mayor precisión, al obtener un resultado decimal, se debe redondear al decimal inmediato superior, utilizando un máximo de dos decimales en la definición de los intervalos de clase

Por ejemplo, si se tienen los siguientes datos:

$\mathbf{X_{max}}$ **= 40.03**, $\mathbf{X_{min}}$ **= 18.73** y **k = 7** ,

entonces

$$\mathbf{X_{max} - X_{min} = 40.03 - 18.73 = 21.3}$$

$$h = \frac{40.03 - 18.73}{7} = \frac{21.3}{7} = 3.0428 \approx 3$$

Si seguimos los criterios tradicionales de redondeo y seleccionamos el entero más cercano, la amplitud de clase (h) sería igual a 3.

Paro si definimos las clases con este valor, tendríamos:

Tabla 4: Intervalos de clase, caso 1

Clase	Intervalo de clase
1	18.73 a 21.73
2	21.73 a 24.73
3	24.73 a 27.73
4	27.73 a 30.73
5	30.73 a 33.73
6	33.73 a 36.73
7	36.73 a 39.73

Como se puede apreciar, la última clase dejaría fuera el valor máximo de 40.03. Es esta la razón por la que los redondeos siempre se realizan al entero o decimal inmediato superior.

Caso 2:

Repitamos el proceso redondeando hacia el entero inmediato superior:

$$h = \frac{40.03 - 18.73}{7} = \frac{21.3}{7} = 3.0428 \approx 4$$

Con lo que los intervalos de clase quedarían:

Tabla 5: Intervalos de clase, caso 2

Clase	Intervalo de clase
1	18.73 a 22.73
2	22.73 a 26.73
3	26.73 a 30.73
4	30.73 a 34.73
5	34.73 a 38.73
6	38.73 a 42.73
7	42.73 a 46.73

En este caso, la última clase excede por completo el valor máximo del conjunto, esto se debe que el valor decimal está mucho más cercano al entero inferior que al superior. Es preferible redondear a dos decimales, considerando la décima y centésima más cercanos al valor obtenido, en este ejemplo tendríamos:

$$h = \frac{40.03 - 18.73}{7} = \frac{21.3}{7} = 3.0428 \approx 3.10 \approx 3.1$$

En este caso, nuestras clases quedarían:

Tabla 6: Intervalos de clase, caso 2a

Clase	Intervalo de clase
1	18.73 – 21.83
2	21.83 – 24.93
3	24.93 – 28.03
4	28.03 – 31.13
5	31.13 – 34.23
6	34.23 – 37.33
7	37.33 – 40.43

Como puede notar, este arreglo de clases distribuye de mejor manera los valores en el número de clases y asegura que el valor máximo está contenido en la última clase.

- Así armamos los intervalos de la siguiente manera:

$a_1 = x_{min}$ $\quad a_2 = a_1 + h \rightarrow [a_1, a_2)$

$a_2 \quad a_3 = a_2 + h \rightarrow [a_2, a_3)$

$a_3 \quad a_4 = a_3 + h \rightarrow [a_3, a_4)$

...

$a_k \quad a_{k+1} = a_k + h \rightarrow [a_k, a_{k+1})$

Y así hasta obtener los intervalos. Observe que al obtener el K intervalo, $[a_k, a_{k+1})$, el valor a_{k+1} debe ser mayor, necesariamente, al valor x_{max}.

Una vez construidos los intervalos, los datos observados se condensarán en cada intervalo de clase que le corresponda y el punto medio de cada clase se le denomina ***marca de clase o centro de clase,*** denotada por MC_k, la cual matemáticamente se obtiene resolviendo la ecuación:

$$MC_k = \frac{a_k + a_{k+1}}{2}$$

Dónde:

K= k-ésimo número de intervalo de clase

La MC siempre es el promedio del límite inferior y el límite superior de la clase.

Una vez determinado el número de clases y el intervalo de clase, procedemos a determinar las frecuencias absolutas de cada una, contando el número de datos que se localizan entre los límites establecidos.

Pongamos un ejemplo:

Un nutriólogo registra el tiempo en que sus pacientes en tratamiento de control de peso pierden 2 Kg de peso, con base en control de dieta y ejercicio físico. Desea conocer el número de días que pasan para que al menos 19 de sus pacientes reduzcan su peso en 2kg, para lo cual, cuenta con la siguiente tabla

Tabla 7: Intervalos para pérdida de 2 Kg de peso. Frecuencia absoluta.

Intervalo de clase (# días)	Frecuencia absoluta
[1 - 3)	1
[3 - 5)	8
[5 - 7)	10
[7 - 9]	9

De acuerdo con la tabla, los intervalos de clase representan el número de días entre los que están los pacientes observados que bajaron 2kg. Por ejemplo, entre 3 a 5 días existen 8 pacientes que bajaron los 2kg, entre 7 a 9 días existen 9 pacientes que bajaron los 2 kg. El interés es verificar en cuantos días, 19 de sus pacientes, bajaron los 2kg y para ello necesitamos sumar la frecuencia de esa misma clase, y de las frecuencias que la preceden, esto es, obtener la ***frecuencia acumulada***.

Determinemos las marcas de clase, para ello utilizaremos la fórmula:

$$MC_k = \frac{a_k + a_{k+1}}{2}$$

Tomando los límites inferior y superior de cada clase, realizamos la ecuación para cada clase:

Tabla 8: Intervalos para pérdida de 2 Kg de peso. Determinación de marcas de clase

Intervalo de clase (# días)	Frecuencia absoluta	Marca de clase
[1 - 3)	1	$\frac{1+3}{2} = 2$
[3 - 5)	8	$\frac{3+5}{2} = 4$
[5 - 7)	10	$\frac{5+7}{2} = 6$
[7 - 9]	9	$\frac{7+9}{2} = 8$

Obtengamos la frecuencia acumulada de nuestro ejemplo.

Tabla 9: Intervalos para pérdida de 2 Kg de peso. Frecuencia acumulada

Intervalo de clase (# días)	Marca de clase	Frecuencia absoluta	Frecuencia acumulada
[1, 3)	2	1	1
[3, 5)	4	8	1+8=9
[5, 7)	6	10	1+8+10=19
[7, 9]	8	9	1+8+10+9=28

De acuerdo con la tabla anterior, podemos determinar que la pérdida de 2 kg de peso para 19 pacientes se obtuvo a entre 5 a 7 días, bajo las condiciones en que se dio el estudio. También es posible afirmar que en promedio los 19 pacientes perdieron 2 kg en 6 días, que es la marca de clase que corresponde al intervalo que contiene 19 en la frecuencia acumulada.

EJEMPLO:

Se obtuvo la muestra de 50 adultos mexicanos encuestados, $n = 50$. En la siguiente tabla se muestra edad en la que fueron diagnosticados con diabetes:

Tabla 10: Edad de diagnóstico en pacientes con diabetes

48	48	79	40	31
52	67	34	21	68
50	49	63	30	53
38	39	35	84	60
52	38	78	36	63
39	50	42	46	51
72	48	50	43	20
50	40	47	47	53
51	71	40	49	50
15	42	37	77	45

Para dar tratamiento a estos datos construiremos primero la "tabla de frecuencias agrupadas".

- *Los valores máximo y mínimo de la muestra son:* $x_{min} = 15 \quad y \quad x_{max} = 84$
- *El valor del rango para este conjunto de datos está dado por:*
- $Rango = X_{max} - X_{min} = 84 - 15 = 69$
- *Aplicando la regla de Sturges se determina el número total de intervalos que se consideraran*
- $k = 1 + 3.322\ log\ (n)$
- *para este caso,* $n = 50$, *entonces*
- $k = 1 + 3.322 * log\ (50) = 6.61$
- *Dado que los decimales se encuentran más cerca del número entero superior que del inferior, redondeamos al entero SUPERIOR más cercano,*
- $k = 7$
- *Determinado la amplitud de cada una de las 7 clases que se han de considerar tenemos:*
- $h = \frac{x_{max} - x_{min}}{k} = \frac{84-15}{7} = \frac{69}{7} = 9.8571$

- *Nuevamente el valor fraccionario está próximo al entero inmediato superior, por lo que se redondea hacia ese valor:*
- $h = \frac{84-15}{7} = \frac{69}{7} = 9.8571 \approx 10.$

La construcción de los 7 intervalos se muestra a continuación:

- $a_1 = x_{min} = 15,$
- $a_2 = a_1 + amplitud = 15 + 10 = 25 \rightarrow [a_1 = 15, a_2 = 25)$
- $a_2 = 25,$
- $a_3 = a_2 + amplitud = 25 + 10 = 35 \rightarrow [a_2 = 25,\ a_3 = 35)$
- $a_3 = 35,$
- $a_4 = a_3 + amplitud = 35 + 10 = 45 \rightarrow [a_3 = 35,\ a_4 = 45)$
- $a_4 = 45,$
- $a_5 = a_4 + amplitud = 45 + 10 = 55 \rightarrow [a_4 = 45,\ a_5 = 55)$
- $a_5 = 55,$
- $a_6 = a_5 + amplitud = 55 + 10 = 65 \rightarrow [a_5 = 55,\ a_6 = 65)$
- $a_6 = 65,$
- $a_7 = a_6 + amplitud = 65 + 10 = 75 \rightarrow [a_6 = 65,\ a_7 = 75)$
- $a_7 = 75,$
- $a_8 = a_7 + amplitud = 75 + 10 = 85 \rightarrow [a_7 = 75,\ a_8 = 85]$

Para validar si la amplitud de clase es adecuada, verificamos que el valor máximo esté contenido en la última clase del arreglo, esto es:

- $X_{max} = 84 \in Clase\ [75 - 85)$

Lo anterior se lee 84 pertenece a la clase [75 – 85). Por lo tanto, la amplitud de clase resulta pertinente.

Dados los intervalos de clase, se procede a contar los datos que pertenecen a cada intervalo. Por ejemplo, para el intervalo $[15, 25)$, buscamos en la muestra cuantos datos pertenecen a ese intervalo.

La frecuencia para el intervalo $[15, 25)$ es 3, para el intervalo [35, 45) la frecuencia es 13.

De manera análoga para los demás intervalos de clase.

Así, tenemos la siguiente tabla de frecuencias agrupadas:

Tabla 11: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia absoluta

Clase	Intervalos de clase	Marca de clase	Frecuencia
1	[15, 25)	20	3
2	[25, 35)	30	3
3	[35, 45)	40	13
4	[45, 55)	50	20
5	[55, 65)	60	3
6	[65, 75)	70	4
7	[75, 85)	80	4
		Total:	50

Nota que la mayor frecuencia de datos se tiene entre los 45 y 54 años, es decir, a esa edad fueron diagnosticados más casos de diabetes.

Obtengamos la frecuencia relativa, recuerda que la expresión matemática es la siguiente:

$$fr_k = \frac{f_k}{n}$$

o en su forma porcentual

$$fr_k = \frac{f_k}{n} \times 100\ \%$$

Tabla 12: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia relativa

Clase	Intervalos de clase	Marca de clase	Frecuencia	Frecuencia relativa
1	[15, 25)	20	3	6%
2	[25, 35)	30	3	6%
3	[35, 45)	40	13	26%
4	[45, 55)	50	20	40%
5	[55, 65)	60	3	6%
6	[65, 75)	70	4	8%
7	[75, 85)	80	4	8%
		Total:	50	100%

La frecuencia acumulada se puede calcular de dos formas; la primera, sumando las frecuencias absolutas de la clase actual y las que la preceden:

Tabla 13: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia acumulada

Clase	Intervalos de clase	Marca de clase	Frecuencia	Frecuencia relativa	Frecuencia acumulada

1	[15, 25)	20	3	6%	3
2	[25, 35)	30	3	6%	3+3=6
3	[35, 45)	40	13	26%	13+3+3=19
4	[45, 55)	50	20	40%	20+13+3+3=39
5	[55, 65)	60	3	6%	3+20+13+3+3=42
6	[65, 75)	70	4	8%	4+3+20+13+3+3=46
7	[75, 85)	80	4	8%	4+3+20+13+3+3=50
		Total:	50	100%	

La segunda forma es sumando la frecuencia absoluta de la clase actual más la frecuencia acumulada de la clase anterior:

Tabla 14: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia acumulada

Clase	Intervalos de clase	Marca de clase	Frecuencia	Frecuencia relativa	Frecuencia acumulada
1	[15, 25)	20	3	6%	3
2	[25, 35)	30	3	6%	3+3=6
3	[35, 45)	40	13	26%	13+6=19
4	[45, 55)	50	20	40%	20+19=39
5	[55, 65)	60	3	6%	3+39=42
6	[65, 75)	70	4	8%	4+42=46
7	[75, 85)	80	4	8%	4+46=50
		Total:	50	100%	

La frecuencia acumulada nos permite conocer los casos o eventos en los que se presenta la característica que estamos analizando, en orden creciente y considerando las clases en orden progresivo; esto es, si seleccionamos el segundo renglón de la tabla anterior, notaremos que tiene un valor de 6 (3 que presentan casos de diabetes en la clase de 15 a 25 años y 3 que presentan casos en la clase de 25 a 35 años); si deseamos estimar los casos de diabetes que se presentan a la edad de 65 años o menos, debemos consultar la frecuencia acumulada de la clase 5 cuyo intervalo va de 55 a 65 años, que para el ejemplo anterior tiene una frecuencia acumulada igual a 42 se debe entender que se tienen 42 casos donde se diagnosticó a un paciente con diabetes y cuya edad oscilaban entre los 15 años (cumplidos) y 65 años (antes de cumplirlos).

Note que cuando se dice "cumplidos", significa que 15 años se incluye en el intervalo; mientras que al decir "antes de cumplirlos", indica que 65 años queda excluido en el intervalo.

Cuando agregamos la siguiente columna, FRECUENCIA RELATIVA ACUMULADA, tenemos la siguiente tabla:

Tabla 15: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia relativa acumulada.

Clase	Intervalo	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	[15, 25)	3	6 %	3	6 %
2	[25, 35)	3	6 %	6	12 %
3	[35, 45)	13	26 %	19	38 %
4	[45, 55)	20	40 %	39	78 %
5	[55, 65)	3	6 %	42	84 %
6	[65, 75)	4	8 %	46	92 %
7	[75, 85)	4	8 %	50	100 %
	Total	50			

Como se puede apreciar, en la columna de FRECUENCIA RELATIVA ACUMULADA es posible consultar los porcentajes acumulados de cada una de las clases considerando el valor de las anteriores, de aquí podemos conocer que 84 % de los casos que presentaron diabetes en el estudio se presentaron en participantes con menos de 65 años.

La importancia de las tablas de frecuencia, y en específico las frecuencias relativas, nos permiten realizar inferencias respecto de la población de la que proceden, por ejemplo, si la tabla anterior hubiese sido obtenida de una población de 10,000 habitantes con casos de diabetes, estos resultados nos permitirían estimar que de esta población, aproximadamente 8,400 tienen 65 años o menos.

La frecuencia relativa es especialmente útil cuando se trata de proyectar los resultados obtenidos en el análisis de una muestra hacia el total de la población. Si tomamos el ejemplo la población de 10,000 habitantes con casos de diabetes, podemos estimar la distribución esperada dentro de la población, de ahí que podamos proponer:

Tabla 16: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Frecuencia relativa acumulada

Clase	Intervalo	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	[15, 25)	600	6 %	600	6 %
2	[25, 35)	600	6 %	1,200	12 %
3	[35, 45)	2,600	26 %	3,800	38 %
4	[45, 55)	4,000	40 %	7,800	78 %
5	[55, 65)	600	6 %	8,400	84 %
6	[65, 75)	800	8 %	9,200	92 %
7	[75, 85)	800	8 %	10,000	100 %
	Total	10,000	100%		

Nota que los porcentajes de frecuencia relativa nos permiten obtener los valores esperados de casos en la población, si conocemos el tamaño de la población y la muestra resulta SER REPRESENTATIVA, esto es, si la muestra tiene un tamaño adecuado y refleja lo mejor posible las características de la población, la distribución esperada será muy cercana a la realidad.

Recordemos que la estadística NO ES UNA CIENCIA EXACTA, y que sus estimaciones pueden diferir en cierta medida de los datos reales, pero permite tener un marco de referencia cercano a la realidad que permita entender el fenómeno.

Cuando trabajamos con clases, es conveniente tener un punto de referencia, el llamado CENTRO DE CLASE o MARCA DE CLASE, que se ubica exactamente al centro del intervalo de clase (h).

Para obtenerla marca de clase, se procede a calcular el promedio entre el límite inferior y el límite superior de cada clase, la fórmula a utilizar es:

$$MC_k = \frac{a_k + a_{k+1}}{2},$$

donde

k= k-ésimo número intervalo de clase.

Calculando los centros de clase, tenemos:

k=1

- $[a_1 = 15, a_2 = 25) \quad \rightarrow \quad MC_1 = \frac{15+25}{2} = 20$

k=2

- $[a_2 = 25, a_3 = 35) \quad \rightarrow \quad MC_2 = \frac{25+35}{2} = 30$

k=3

- $[a_3 = 35, a_4 = 45) \quad \rightarrow \quad MC_3 = \frac{35+45}{2} = 40$

k=4

- $[a_4 = 45, a_5 = 55) \quad \rightarrow \quad MC_4 = \frac{45+55}{2} = 50$

k=5

- $[a_5 = 55, a_6 = 65) \quad \rightarrow \quad MC_5 = \frac{55+65}{2} = 60$

k=6

- $[a_6 = 65, a_7 = 75) \quad \rightarrow \quad MC_6 = \frac{65+75}{2} = 70$

k=7

- $[a_7 = 75, a_8 = 85] \quad \rightarrow \quad MC_7 = \frac{75+85}{2} = 80$

Así, se tiene la siguiente tabla de frecuencias agrupadas:

Tabla 17: Tabla de frecuencias agrupadas de edad de diagnóstico en pacientes con diabetes. Marca de clase.

Clase	Intervalo	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada	Marca de clase
1	[15, 25)	600	6 %	600	6 %	20
2	[25, 35)	600	6 %	1,200	12 %	30
3	[35, 45)	2,600	26 %	3,800	38 %	40
4	[45, 55)	4,000	40 %	7,800	78 %	50
5	[55, 65)	600	6 %	8,400	84 %	60
6	[65, 75)	800	8 %	9,200	92 %	70
7	[75, 85)	800	8 %	10,000	100 %	80
	Total	10,000	100 %			

Cuando estamos analizando datos provenientes de un estudio que involucra grandes cantidades de participantes (un censo poblacional, por ejemplo), normalmente la información se nos presenta resumida en tablas, de forma que tenemos poco o nulo acceso a los datos explícitos obtenidos en el estudio. Por ejemplo, si consultamos los datos de población por grupo de edad de alguna ciudad, obtendremos una tabla como la siguiente:

Tabla 18: Tabla de frecuencia por grupos de edad de la población, Ciudad X

Grupo de edad	Frecuencia absoluta	Centro de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
0 - 5	43,969	2.5	4.2	43,969	4.2
5 - 10	48,040	7.5	4.59	92,009	8.79
10 - 15	53,102	12.5	5.07	145,111	13.86
15 - 20	72,077	17.5	6.89	217,188	20.75
20 - 25	88,006	22.5	8.41	305,194	29.15
25 - 30	84,858	27.5	8.11	390,052	37.26
30 - 35	84,440	32.5	8.07	474,492	45.33
35 - 40	78,868	37.5	7.53	553,360	52.86
40 - 45	72,634	42.5	6.94	625,994	59.8
45 - 50	69,840	47.5	6.67	695,834	66.47
50 - 55	65,948	52.5	6.3	761,782	72.77
55 - 60	50,673	57.5	4.84	812,455	77.61
60 - 65	57,869	62.5	5.53	870,324	83.14
65 - 70	54,981	67.5	5.25	925,305	88.39
70 - 75	46,981	72.5	4.49	972,286	92.88
75 - 80	74,526	77.5	7.12	1,046,812	100
Población total	1,046,812				

En esta tabla se encuentra resumida la información procedente de la población, pero difícilmente tendremos acceso a los 1,046,812 registros de los que se generó la tabla y aún si lo tuviéramos, el tiempo invertido en procesar esa información sería muy grande. En estos casos, es posible obtener las medidas de tendencia central y de dispersión a partir de los datos agrupados mostrados en la tabla.

MEDIDAS DE TENDENCIA CENTRAL PARA DATOS AGRUPADOS

Las medidas de tendencia central que abordaremos cuando se tiene una serie de datos agrupados son las mismas que se trabajaron en el tema anterior (para datos no agrupados), es decir, la media aritmética, la mediana y la moda.

Recuerda que las medidas de tendencia central te ayudan a tener un parámetro que te da información sobre el centro de distribución de la muestra que se analiza.

MEDIA DE DATOS AGRUPADOS

La **media aritmética** se calcula con la siguiente expresión matemática:

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

Dónde:

$\bar{X}$= Media

M C_i = Centro de clase de la clase i

F_i = Frecuencia absoluta de la Clase i

N = Número total de elementos

k = k-ésimo numero de intervalo de clase

Para mayor facilidad todos los cálculos se verán reflejados mediante la tabla 1, para la media necesitamos la multiplicación de la marca de clase de cada intervalo por la frecuencia de cada uno. Como se muestra a continuación:

Tabla 19: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la media.

Grupo de edad	Frecuencia absoluta	Centro de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa	M C * Fa
0 - 5	43,969	2.5	4.2	43,969	4.2	109,922.5
5 - 10	48,040	7.5	4.59	92,009	8.79	360,300
10 - 15	53,102	12.5	5.07	145,111	13.86	663,775
15 - 20	72,077	17.5	6.89	217,188	20.75	1,261,347.5
20 - 25	88,006	22.5	8.41	305,194	29.15	1,980,135
25 - 30	84,858	27.5	8.11	390,052	37.26	2,333,595
30 - 35	84,440	32.5	8.07	474,492	45.33	2,744,300
35 - 40	78,868	37.5	7.53	553,360	52.86	2,957,550
40 - 45	72,634	42.5	6.94	625,994	59.8	3,086,945
45 - 50	69,840	47.5	6.67	695,834	66.47	3,317,400
50 - 55	65,948	52.5	6.3	761,782	72.77	3,462,270
55 - 60	50,673	57.5	4.84	812,455	77.61	2,913,697.5

Grupo de edad	Frecuencia absoluta	Centro de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa	M C * Fa
60 - 65	57,869	62.5	5.53	870,324	83.14	3,616,812.5
65 - 70	54,981	67.5	5.25	925,305	88.39	3,711,217.5
70 - 75	46,981	72.5	4.49	972,286	92.88	3,406,122.5
75 - 80	74,526	77.5	7.12	1,046,812	100	5,775,765
Población total	1,046,812		100	Total Mc *fa		41,701,155

De acuerdo con lo obtenido

$\sum_{i=1}^{7} MC_i f_i = 41{,}701{,}155$

y sabemos por la tabla que n = 1,046,812

Sustituyendo en la ecuación, tenemos

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} MC_i f_i, = \frac{41{,}701{,}155}{1{,}046{,}812} = 39.84$$

De donde podemos afirmar que la media de la población estudiada es igual a 39.84 años, sin necesidad de realizar la suma directa de más de un millón de datos que se recabaron para la construcción de la tabla. Note que el centro de clase es un elemento de suma importancia para el cálculo de la media de datos agrupados.

Repitamos el mismo proceso con un ejemplo más simple para reafirmar el proceso:

Tabla 20: Tabla de frecuencia para cálculo de media aritmética de datos agrupados

# Intervalos	Intervalos de clase	Frecuencia	Fr	Fa	MC	MC*f
1	[15, 25]	3	6%	3	20	60
2	[25, 35]	3	6%	6	30	90
3	[35, 45]	13	26%	19	40	520
4	[45, 55]	20	40%	39	50	1000
5	[55, 65]	3	6%	42	60	180
6	[65, 75]	4	8%	46	70	280
7	[75, 85]	4	8%	50	80	320
	Total	50	100%		Total	2450

Nota que $\sum_{i=1}^{7} MC_i f_i = 2\,450$, sustituyendo en la expresión

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{7} MC_i f_i = \frac{1}{50}(2\,450) = 49$$

En promedio a la edad de 49 años le detectaron diabetes a la muestra de los 50 habitantes encuestados.

MEDIANA DE DATOS AGRUPADOS

De la misma forma que calculamos la media aritmética para datos agrupados, podemos obtener el valor del dato que ocupa exactamente el centro de la distribución de valores, aun cuando esto se encuentren agrupados en clases.

Antes de dar paso al cálculo, definamos para cada intervalo de clase $[a_k, a_{k+1}]$ lo siguiente:

- $L = a_k$, *es el límite inferior de la clase*
- $U = a_{k+1}$ *es el límite superior de la clase.*
- $R_k = U - L$ *es el rango, y se expresa como la diferencia entre el límite superior y el límite inferior de la clase.*

La **mediana** se calcula con la siguiente fórmula matemática:

$$Me = L_k + \frac{\frac{n}{2} - fa_{k-1}}{f_k} \times R_k$$

A continuación, con base en el ejemplo dado previamente, se mostrará como calcular cada elemento de la expresión matemática y así determinar la mediana. En este caso iniciaremos con la tabla más sencilla y concluiremos con la más complicada, a fin de que el proceso se aprecie claramente con un menos número de datos.

Para obtener la mediana de datos agrupados debemos seguir los pasos que se mencionan a continuación:

- *Identifica la clase que contiene la mediana. Para esto se debe de buscar la clase cuyo intervalo donde se encuentre el valor de* $\frac{n}{2}$ *o el 50 % de la frecuencia relativa acumulada, en caso de que el valor no se encuentre dentro de la frecuencia acumulada, se busca el entero inmediato superior; es más fácil considerar la frecuencia relativa acumulada pues si no aparece el dato explícito de 50 % se toma el valor porcentual superior más cercano, en el ejemplo corresponde a la clase cuyo intervalo es [45, 55), pues tiene una frecuencia relativa acumulada de 78 %, que es la frecuencia relativa acumulada superior más cercana a 50 %..*

Sea $\frac{n}{2} = \frac{50}{2} = 25$,

- *Buscamos la posición 25 en la columna de la frecuencia acumulada, como en la tabla no tenemos exactamente el valor 25 escogemos el mayor entero próximo, en nuestro caso sombreamos la clase de la mediana:*

Tabla 21: Tabla de frecuencia para cálculo de mediana de datos agrupados

Clase	Intervalo	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	[15, 25)	3	6 %	3	6 %
2	[25, 35)	3	6 %	6	12 %
3	[35, 45)	13	26 %	19	38 %
4	[45, 55)	20	40 %	39	78 %
5	[55, 65)	3	6 %	42	84 %
6	[65, 75)	4	8 %	46	92 %
7	[75, 85)	4	8 %	50	100 %
	Total	50			

Como podrá notar, corresponde exactamente a la clase con la frecuencia relativa acumulada superior más cercana al 50 %; puedes elegir el método que se te facilite más, ambos son correctos.

Una vez ubicada la clase que contiene a la mediana (la cual divide a la distribución exactamente en 50 %), esto es, la clase $k = 4$ con el intervalo de clase $[45, 55)$, obtengamos los elementos necesarios para el cálculo de la mediana.

- L_k, *será el límite inferior del intervalo de clase* k.

En nuestro caso, el límite inferior del intervalo [45, 55) es $L_4 = 45$,

- fa_{k-1}, *que es la frecuencia acumulada anterior a la clase de la mediana.*

En nuestro caso, k=4, entonces $fa_{4-1} = 19$.

- f_k, *corresponde a la frecuencia donde se identificó la clase de la mediana.*

En nuestro caso, k=4, entonces la frecuencia $f_4 = 20$

- R_k, *será el rango del intervalo de clase donde está la clase de la mediana.*

$R_4 = U_4 - L_4 = 55 - 45 = 10$, sustituyendo en la fórmula de la mediana

$$Me = L_4 + \frac{\frac{n}{2} - fa_3}{f_4} \times R_4 = 45 + \frac{25 - 19}{20} \times 10 = 45 + \frac{6}{20} \times 10 = 48$$

Por lo tanto, la edad que divide al conjunto de datos en dos partes iguales corresponde a la $\boldsymbol{Me = 48}$.

Repitamos el proceso para la tabla de distribución de población por grupo de edad:

Tabla 22: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la mediana.

Clase	Grupo de edad	Frecuencia absoluta	Centro de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	0 - 5	43,969	2.5	4.2	43,969	4.2
2	5 - 10	48,040	7.5	4.59	92,009	8.79
3	10 - 15	53,102	12.5	5.07	145,111	13.86
4	15 - 20	72,077	17.5	6.89	217,188	20.75
5	20 - 25	88,006	22.5	8.41	305,194	29.15
6	25 - 30	84,858	27.5	8.11	390,052	37.26
7	30 - 35	84,440	32.5	8.07	474,492	45.33
8	35 - 40	78,868	37.5	7.53	553,360	52.86
9	40 - 45	72,634	42.5	6.94	625,994	59.8
10	45 - 50	69,840	47.5	6.67	695,834	66.47
11	50 - 55	65,948	52.5	6.3	761,782	72.77
12	55 - 60	50,673	57.5	4.84	812,455	77.61
13	60 - 65	57,869	62.5	5.53	870,324	83.14
14	65 - 70	54,981	67.5	5.25	925,305	88.39
15	70 - 75	46,981	72.5	4.49	972,286	92.88
16	75 - 80	74,526	77.5	7.12	1,046,812	100
Población total		1,046,812		100		

1) *Localizar la clase que contiene a la mediana: en este caso corresponde a la clase 8 cuyo intervalo va de 35 a 40 años, ya que tiene una frecuencia relativa acumulada igual a 52.86 %*
2) *El límite inferior de esta clase es* $L_8 = 35$
3) *El límite superior de esta clase* $U_8 = 40$
4) *La frecuencia acumulada de la clase anterior a la clase que contiene la mediana, en este caso la Clase 7 con intervalo de 30 a 35 años, que corresponde a 474,492*
5) *La frecuencia absoluta de la clase 8, que es la que contiene a la mediana, en este caso 78,868*

6) El Rango del intervalo que contiene a la mediana, que está dado por:

- $R_8 = U_8 - L_8 = 40 - 35 = 5$
- *Este rango, si se trata de clases con el mismo tamaño, es el mismo para todas las clases.*

A continuación, procedemos a la fórmula para el cálculo de la mediana para datos agrupados:

$$Me = L_4 + \frac{\frac{n}{2} - fa_3}{f_4} \times R_4 = 35 + \frac{\frac{1046812}{2} - 474{,}492}{78{,}868} \times 5$$
$$= 35 + \frac{523{,}406 - 474{,}492}{78{,}868} \times 5$$
$$= 35 + 3.10 = 38.10$$

La mediana para este conjunto de datos es 38.10 años.

MODA DE DATOS AGRUPADOS

La **moda (el dato que más se repite)** se obtiene mediante la siguiente formula:

$$Mo = L_k + \frac{f_k - f_{k-1}}{(f_k - f_{k-1}) + (f_k - f_{k+1})} \times R_k$$

Dónde:

Mo = Moda

Lk = Límite inferior de la clase

F = Frecuencia absoluta

K = clase con mayor frecuencia absoluta

A continuación, siguiendo con el ejemplo, se mostrará como calcular cada elemento de la expresión matemática para así determinar la moda.

- *Identifica la clase modal. Para esto se busca el intervalo de clase con la frecuencia más alta.*

Para la moda, sombreamos en la tabla el intervalo de clase con mayor frecuencia,

Tabla 23: Tabla de frecuencia acumulada. Cálculo de la moda.

# Intervalos	Intervalos de clase	Frecuencia	Fr	Fa
1	[15, 25]	3	6%	3
2	[25, 35]	3	6%	6
3	[35, 45]	13	26%	19
4	[45, 55]	20	40%	39
5	[55, 65]	3	6%	42
6	[65, 75]	4	8%	46
7	[75, 85]	4	8%	50
	Total	50	100%	

corresponde al intervalo $[45, 55)$ posicionándonos en el número de intervalo $k = 4$.

- *L_k, será el límite inferior del intervalo de clase k.*

En nuestro caso, el límite inferior del intervalo [45, 55) es $L_4 = 45$,

- *f_k, corresponde a la frecuencia del intervalo de clase k, clase modal.*

Si k=4, entonces $f_4 = 20$

- *f_{k-1}, corresponde a la frecuencia anterior del intervalo de clase k, clase modal.*

Si k=4, entonces $f_{4-1} = 13$

- *f_{k+1}, corresponde a la frecuencia posterior del intervalo de clase k, clase modal.*

Si k=4, entonces $f_{4+1} = 3$ y

Nota. Si la clase modal corresponde al primer intervalo, entonces $f_{k-1} = 0$. Si la clase modal está en el último intervalo, entonces $f_{k+1} = 0$.

- *R_k, será el rango del intervalo de clase donde está la clase de la mediana.*

$$R_4 = U_4 - L_4 = 55 - 45 = 10$$

De la misma manera que en el caso anterior, SI LAS CLASES SON DEL MISMO TAMAÑO, EL RANGO SERÁ EL MISMO PARA TODAS LAS CLASES.

sustituyendo en la fórmula de la moda

$$Mo = L_4 + \frac{f_4 - f_3}{(f_4 - f_3) + (f_4 - f_5)} \times R_4$$

$$Mo = 45 + \frac{20 - 13}{(20 - 13) + (20 - 3)} \times 10 = 45 + \frac{7}{24} \times 10$$
$$= 47.92$$

Por lo que EL VALOR DE LA MODA, considerando una distribución UNIMODAL, correspondiente a 47.92.

Sabemos que se trata de una distribución con una sola moda debido a que únicamente una clase es la que tiene la FRECUENCIA ABSOLUTA más alta, si existieran don clases con el mismo valor de frecuencia absoluta y estos fueran los más altos, tendremos que calcular la moda para cada caso y asumiríamos una distribución BIMODAL, si hubiera tres clases con la frecuencia absoluta más alta y de igual valor, se calcularían tres modas y asumiríamos una distribución TRIMODAL y así sucesivamente.

Calculemos ahora la moda para la tabla:

Tabla 24: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la moda.

Clase	Grupo de edad	Frecuencia absoluta	Centro de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
1	0 – 5	43969	2.5	4.2	43969	4.2
2	5 – 10	48040	7.5	4.59	92009	8.79
3	10 - 15	53102	12.5	5.07	145111	13.86
4	15 - 20	72077	17.5	6.89	217188	20.75
5	20 - 25	88006	22.5	8.41	305194	29.15
6	25 - 30	84858	27.5	8.11	390052	37.26
7	30 - 35	84440	32.5	8.07	474492	45.33
8	35 - 40	78868	37.5	7.53	553360	52.86
9	40 - 45	72634	42.5	6.94	625994	59.8
10	45 - 50	69840	47.5	6.67	695834	66.47
11	50 - 55	65948	52.5	6.3	761782	72.77
12	55 - 60	50673	57.5	4.84	812455	77.61
13	60 - 65	57869	62.5	5.53	870324	83.14
14	65 - 70	54981	67.5	5.25	925305	88.39
15	70 - 75	46981	72.5	4.49	972286	92.88
16	75 - 80	74526	77.5	7.12	1046812	100
Población total		1046812		100		

Sigamos el algoritmo mencionado:

1) *Identifiquemos si existe una o varias clases modales con el mismo valor de frecuencia absoluta, en este caso la clase modal solo es una y corresponde a la Clase 5, con intervalo 20 – 25.*
2) *Tomemos el valor del límite inferior de la clase que contiene la moda (o clase modal), en este caso tenemos que $L_5 = 20$*
3) *Tomemos el valor de la frecuencia absoluta de la clase muestral, en este caso tenemos que $F_5 = 88{,}006$*
4) *Tomemos la frecuencia absoluta de la clase anterior a la clase modal, en este caso se trata de la clase (5-1), o sea Clase 4, con intervalo de 15 a 20, lo que nos devuelve un valor de 72,077.*
5) *Tomemos la frecuencia absoluta de La clase posterior a la clase modal, en este caso se trata de la clase (5 + 1), o sea la Clase 6, con intervalo de 25 a 30, lo que nos da un valor de 84,858.*
   - *Como la clase modal no es ni la primera ni la última, proseguimos con el algoritmo.*
6) *Finalmente, tomamos el valor del rango, dado por:*
   - $R_5 = U_5 - L_5 = 25 - 20 = 5$

Ahora, sustituyamos valores en la fórmula:

$$Mo = L_5 + \frac{f_5 - f_4}{(f_5 - f_4) + (f_5 - f_6)} \times R_5$$

$$= 20 + \frac{88{,}006 - 72{,}077}{(88{,}006 - 72{,}077) + (88{,}006 - 84{,}858)} \times 5$$

$$Mo = 20 + \frac{88{,}006 - 72{,}077}{(88{,}006 - 72{,}077) + (88{,}006 - 84{,}858)} \times 5$$

$$= 20 + \frac{15{,}929}{(15{,}929) + (3148)} \times 5$$

$$Mo = 20 + \frac{15{,}929}{19077} \times 5 = 20 + (0.8349 \times 5) = 24.175$$

El valor de la moda de datos agrupados es 24.175

Medidas de dispersión de datos agrupados

Recuerda que las medidas de dispersión posibilitan visualizar la variabilidad o dispersión de los datos asociados a una variable y determinar qué tan alejados están de la media. Las medidas de dispersión que abordaremos cuando se tienen datos agrupados son la desviación estándar y la varianza.

La **varianza** se obtiene mediante la expresión:

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

Donde:

S^2 = Varianza

n = Número total de valores en la distribución

f_i = Frecuencia de la clase i

MC_i = Marca o centro de clase de la clase i

I = i-ésimo valor del arreglo

Mientras que la **desviación estándar** se obtiene como la raíz cuadrada de la varianza, matemáticamente se expresa como

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

Al calcular la varianza, obtenemos de inmediato la desviación estándar al calcular la raíz cuadrada del valor obtenido.

Ahora, con base en la tabla 1, obtenemos a partir de la tabla, $MC_k - \bar{x}$, dada la media $\bar{x} = 49$.

Tabla 25: Tabla de frecuencia acumulada. Cálculo de la varianza, paso 1.

Clase	Intervalos de clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulad	Marca de clase (MC)	$MC - \bar{x}$
1	[15, 25)	3	6%	3	20	-29
2	[25, 35)	3	6%	6	30	-19
3	[35, 45)	13	26%	19	40	-9
4	[45, 55)	20	40%	39	50	1
5	[55, 65)	3	6%	42	60	11
6	[65, 75)	4	8%	46	70	21
7	[75, 85]	4	8%	50	80	31
	Total:	50	100%			

Posteriormente $MC_k - \bar{x}$ se eleva al cuadrado y una vez elevado al cuadrado se multiplica por su frecuencia y así obtenemos la suma total.

Tabla 26: Tabla de frecuencia acumulada. Cálculo de la varianza, paso 2.

Clase	Intervalos de clase	Frecuencia absoluta (f_i)	Marca de clase (MC)	$MC - \bar{x}$	$(MC - \bar{x})^2$	$fa * (MC - \bar{x})^2$
1	[15, 25)	3	20	-29	841	2,523
2	[25, 35)	3	30	-19	361	1,083
3	[35, 45)	13	40	-9	81	1,053
4	[45, 55)	20	50	1	1	20
5	[55, 65)	3	60	11	121	363
6	[65, 75)	4	70	21	441	1,764
7	[75, 85]	4	80	31	961	3,844
					Total:	**10,650**

Nota que, al elevar al cuadrado, los signos negativos desaparecen y podemos estimar valores de dispersión efectivos, sin importar el sentido.

Así tenemos la suma $\sum_{i=1}^{7} f_i(MC_i - \bar{x})^2 = 10\,650$ y sustituyendo en la formula de la varianza se obtiene

$$s^2 = \frac{1}{50-1}\sum_{i=1}^{7} f_i(MC_i - \bar{x})^2 = \frac{1}{49}(10\,650) = 217.35$$

Para la desviación estándar, solo es la raíz de la varianza, es decir

$$s^2 = 217.35 \quad \rightarrow \quad \sqrt{s^2} = \sqrt{217.35} \quad \rightarrow \quad \sigma = 14.74$$

Así la desviación estándar es de 14.74.

Ahora probemos con la tabla

Tabla 27: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la varianza, paso 1.

Clase	Grupo de edad	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada	Centro de clase
1	0 - 5	43,969	4.2	43,969	4.2	2.5
2	5 - 10	48,040	4.59	92,009	8.79	7.5
3	10 - 15	53,102	5.07	145,111	13.86	12.5
4	15 - 20	72,077	6.89	217,188	20.75	17.5
5	20 - 25	88,006	8.41	305,194	29.15	22.5
6	25 - 30	84,858	8.11	390,052	37.26	27.5
7	30 - 35	84,440	8.07	474,492	45.33	32.5
8	35 - 40	78,868	7.53	553,360	52.86	37.5
9	40 - 45	72,634	6.94	625,994	59.8	42.5
10	45 - 50	69,840	6.67	695,834	66.47	47.5
11	50 - 55	65,948	6.3	761,782	72.77	52.5
12	55 - 60	50,673	4.84	812,455	77.61	57.5
13	60 - 65	57,869	5.53	870,324	83.14	62.5
14	65 - 70	54,981	5.25	925,305	88.39	67.5
15	70 - 75	46,981	4.49	972,286	92.88	72.5
16	75 - 80	74,526	7.12	1,046,812	100	77.5
	Población total	1,046,812	100			

Antes que nada, hay que notar que estos datos provienen de una POBLACIÓN, por lo que la notación que

utilizaremos para la media es **μ**, que se utiliza para MEDIA POBLACIONAL, y $\mathbf{s^2}$, que se utiliza para VARIANZA POBLACIONAL.

1) *Obtengamos la media: Este valor lo determinamos cuando se revisó el tema de MEDIA DE DATOS AGRUPADOS, el valor de la media para este Conjunto de datos es de 39.84.*
2) *Agreguemos una columna donde calcularemos la diferencia de cada centro de clase menos la media, esto es:*

Tabla 28: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la varianza, paso 2.

Clase	Grupo de edad	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada	Centro de clase	MC - μ
1	0 - 5	43,969	4.2	43,969	4.2	2.5	-37.34
2	5 - 10	48,040	4.59	92,009	8.79	7.5	-32.34
3	10 - 15	53,102	5.07	145,111	13.86	12.5	-27.34
4	15 - 20	72,077	6.89	217,188	20.75	17.5	-22.34
5	20 - 25	88,006	8.41	305,194	29.15	22.5	-17.34
6	25 - 30	84,858	8.11	390,052	37.26	27.5	-12.34
7	30 - 35	84,440	8.07	474,492	45.33	32.5	-7.34
8	35 - 40	78,868	7.53	553,360	52.86	37.5	-2.34
9	40 - 45	72,634	6.94	625,994	59.8	42.5	2.66
10	45 - 50	69,840	6.67	695,834	66.47	47.5	7.66
11	50 - 55	65,948	6.3	761,782	72.77	52.5	12.66
12	55 - 60	50,673	4.84	812,455	77.61	57.5	17.66
13	60 - 65	57,869	5.53	870,324	83.14	62.5	22.66
14	65 - 70	54,981	5.25	925,305	88.39	67.5	27.66
15	70 - 75	46,981	4.49	972,286	92.88	72.5	32.66
16	75 - 80	74,526	7.12	1,046,812	100	77.5	37.66
Población total		1,046,812	100				

3) *Elevemos las diferencias al cuadrado:*

Tabla 29: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la varianza, paso 3.

Clase	Grupo de edad	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa	Centro de clase	MC - μ	$(MC-\mu)^2$
1	0 - 5	43,969	4.2	43,969	4.2	2.5	-37.34	1394.2756
2	5 - 10	48,040	4.59	92,009	8.79	7.5	-32.34	1045.8756
3	10 - 15	53,102	5.07	145,111	13.86	12.5	-27.34	747.4756
4	15 - 20	72,077	6.89	217,188	20.75	17.5	-22.34	499.0756
5	20 - 25	88,006	8.41	305,194	29.15	22.5	-17.34	300.6756
6	25 - 30	84,858	8.11	390,052	37.26	27.5	-12.34	152.2756
7	30 - 35	84,440	8.07	474,492	45.33	32.5	-7.34	53.8756
8	35 - 40	78,868	7.53	553,360	52.86	37.5	-2.34	5.4756
9	40 - 45	72,634	6.94	625,994	59.8	42.5	2.66	7.0756
10	45 - 50	69,840	6.67	695,834	66.47	47.5	7.66	58.6756
11	50 - 55	65,948	6.3	761,782	72.77	52.5	12.66	160.2756
12	55 - 60	50,673	4.84	812,455	77.61	57.5	17.66	311.8756
13	60 - 65	57,869	5.53	870,324	83.14	62.5	22.66	513.4756
14	65 - 70	54,981	5.25	925,305	88.39	67.5	27.66	765.0756
15	70 - 75	46,981	4.49	972,286	92.88	72.5	32.66	1066.6756
16	75 - 80	74,526	7.12	1,046,812	100	77.5	37.66	1418.2756
Población total		1,046,812	100					

4) *Multipliquemos cada resultado por su frecuencia y obtengamos la sumatoria:*

Tabla 29: Tabla de frecuencia por grupos de edad de la población, Ciudad X. Cálculo de la varianza, paso 4.

Clase	Grupo de edad	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada	Centro de clase	MC - μ	(MC-μ)²	f-(MC -μ)²
1	0 - 5	43,969	4.2	43,969	4.2	2.5	-37.34	-37.34	1,394.2756
2	5 - 10	48,040	4.59	92,009	8.79	7.5	7.5	-32.34	1,045.8756
3	10 - 15	53,102	5.07	145,111	13.86	12.5	12.5	-27.34	747.4756
4	15 - 20	72,077	6.89	217,188	20.75	17.5	17.5	-22.34	499.0756
5	20 - 25	88,006	8.41	305,194	29.15	22.5	22.5	-17.34	300.6756
6	25 - 30	84,858	8.11	390,052	37.26	27.5	27.5	-12.34	152.2756
7	30 - 35	84,440	8.07	474,492	45.33	32.5	32.5	-7.34	53.8756
8	35 - 40	78,868	7.53	553,360	52.86	37.5	37.5	-2.34	5.4756
9	40 - 45	72,634	6.94	625,994	59.8	42.5	42.5	2.66	7.0756
10	45 - 50	69,840	6.67	695,834	66.47	47.5	47.5	7.66	58.6756
11	50 - 55	65,948	6.3	761,782	72.77	52.5	52.5	12.66	160.2756
12	55 - 60	50,673	4.84	812,455	77.61	57.5	57.5	17.66	311.8756
13	60 - 65	57,869	5.53	870,324	83.14	62.5	62.5	22.66	513.4756
14	65 - 70	54,981	5.25	925,305	88.39	67.5	67.5	27.66	765.0756
15	70 - 75	46,981	4.49	972,286	92.88	72.5	72.5	32.66	1,066.6756
16	75 - 80	74,526	7.12	1,046,812	100	77.5	77.5	37.66	1,418.2756
		1,046,8	100						490,153,449

El valor total obtenido de

$$\sum_{i=1}^{16} f_i(MC_i - \mu)^2 = 490{,}153{,}449$$

Como estos datos provienen de una POBLACIÓN, dividimos entre **n**; recuerda que si los datos provienen de una muestra, se debe dividir entre $\mathbf{n-1}$. Entonces:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{7} f_i(MC_i - \mu)^2 = \frac{1}{1{,}046{,}812}(490{,}153{,}449) = 468.23$$

$$\sigma^2 = 468.23$$

DESVIACIÓN ESTÁNDAR DE DATOS AGRUPADOS:

Como en el caso de la desviación estándar de datos no agrupados, una vez calculada la varianza, basta con obtener su raíz cuadrada para conocer el valor de la desviación estándar, entonces:

$$\sigma = \sqrt{\sigma^2}$$

Esto es:

$$\sigma = \sqrt{468.23}$$

$$\sigma = 21.63$$

PRÓXIMO NÚMERO DE LA SERIE:

Estadística descriptiva (IV): Presentación de datos

Referencias

Anderson, D. R., SWeeney, D., & Williams, T. A. (1999). Estadística para la Administración y Economía. México DF, México: International Thompson Editores.

Departamento de Didáctica de la Matemática. (2011). Estadística con proyectos. (C. Batanero, & C. Díaz, Eds.) Granada, España: Facultad de Ciencias de la Educación, Universidad de Granada.

García Pérez, A. (2008). Estadística aplicada: conceptos básicos (2a edición ed.). Madrid, España: Educación permanente / Universidad Nacional de Educación a Distancia.

Hanlle, V. (2011). Proyecto de estudio de las pausas activas en el Clima Laboral y su influencia e impacto para la motivación y satisfación física de los empleados de Premex Ecuador en la Ciudad de Quito. Tesis. Universidad de las Américas, Quito.

Jaraiseh, N. (2015). Estrés Laboral y Sindrome de Burnout: Pausas activas como método de afrontamiento. Tesis . Universdad Internacioanl SEK, Quito.

Kazmier, L. J., Díaz Mata, A., & Eslava Gómez, G. (1991). Estadística Aplicada a Administración y Economía. Naucálpan, Estado de MéxicoMéxico, Atlacomulco, México: McGraw Hill.

Pérez López, C. (1999). Control estadístico de la calidad. Madrid, España: Alfa Omega.

Wackerly, D. D., Mendenhall III, W., & Scheaffer, R. (2010). Estadística Matemática con aplicaciones. México, D.F., México: Cengage Learning Editores, S.A.

Declaración de conflicto de intereses

El autor declara no tener ningún interés comercial o asociativo que represente un conflicto de intereses en relación con el trabajo presentado.