Ejemplo de datos agrupados y sus medidas estadísticas

1 Problema: análisis de los ingresos del sector de la construcción en el 2019

1.1 Contexto

En agosto de 2020 el periódico económico colombiano *La República* publicó un artículo acerca de los ingresos generados por el sector de la construcción en el 2019. El artículo revela que las empresas con mayores ventas recibieron ingresos del orden de billones de pesos mientras que las menos posicionadas sus ingresos estuvieron del orden de miles de millones de pesos. La siguiente tabla muestra los ingresos recibidos por 40 empresas del sector de la construcción en el 2019.

0.4	0.7	0.2	0.4	0.2	1.1	0.5	0.5
0.7	0.5	0.5	0.2	0.2	0.7	0.5	0.8
0.8	0.4	0.8	0.4	0.5	0.7	0.7	0.8
0.8	0.5	0.5	0.4	0.5	1.0	0.8	1.0
0.4	0.8	0.8	0.4	0.4	1.0	0.5	0.5

Los ingresos están valorados en billones de pesos.

Vita, Laura (6 de agosto de 2020). *Constructoras generaron ingresos por $22,8 billones pese a decrecimiento del sector*. La República. Recuperado de https://www.larepublica.co

1.2 Preguntas

Se requiere estudiar si los ingresos realmente son bastante altos y cuál es su distribución. Para conseguir tal análisis, realizar:

1. Realizar una tabla de datos agrupados.
2. Encontrar la media y desviación de los ingresos.
3. Representar gráficamente la información recogida.

2 Solución: tratamiento usando datos agrupados

El procedimiento para analizar esta situación se realiza a través del manejo de datos agrupados. Por tanto, el esquema a seguir es:

1. Construir los intervalos de clase.
2. Construir la tabla de distribución de frecuencia con las frecuencias absolutas y las marcas de clase.
3. Calcular medidas estadísticas: la media, la desviación estándar.
4. Representar la información en un histograma.
5. Luego del análisis, emitir conclusiones orientandas hacia la respuesta de la pregunta propuesta.

2.1 Construcción de los intervalos de clase

Se hallan los datos máximo y mínimo:

$$x_{max} = 1.1$$
$$x_{min} = 0.2$$

El rango es por tanto:

$$R = x_{max} - x_{min} = 1.1 - 0.2 = 0.9$$

El número de intervalos para una cantidad de $N = 40$ datos, según la fórmula de Sturges es:

$$k = 1 + \frac{\log N}{\log 2} = 1 + \frac{\log 40}{\log 2} = 1 + \frac{1.60}{0.30} = 6.3 \approx 6$$

El ancho de cada intervalo de clase es:

$$C = \frac{R}{k} = \frac{0.9}{6} = 0.15 \approx 0.2$$

Se ha aproximado a una cifra decimal, en lugar de dos, ya que los datos se presentan con una cifra decimal.

Para construir los intervalos, el límite inferior del **primer intervalo** se toma con una décima menos que el dato mínimo, es decir se toma como límite inferior 0.1. El límite superior se obtiene sumando el ancho de clase al límite inferior, es decir

$$0.1 + 0.2 = 0.3$$

El límite inferior del **segundo intervalo** tendrá una décima más que el límite superior del primer intervalo, o sea

$$0.3 + 0.1 = 0.4$$

a su vez, a este límite se le suma el ancho de clase, luego vale 0.6. De forma similar para los restantes intervalos como muestra la tabla de abajo.

Intervalo	Lím. inferior	Lím. superior
Primero	0.1	0.3
Segundo	0.4	0.6
Tercero	0.7	0.9
Cuarto	1.0	1.2
Quinto	1.3	1.5
Sexto	1.6	1.8

2.2 Tabla de distribución de frecuencia

Determinados los intervalos de clase, se inicia la clasificación de los datos en los respectivos intervalos haciendo los respectivos conteos de frecuencia. A su vez se determinan las marcas de clase a partir de la media (promedio) de los límites del intervalo. Según lo anterior, la tabla de de distribución de frecuencia de la situación queda como sigue:

Ingresos (billones \$)	Frecuencia absoluta (f)	Marca de clase (x)
0.1 - 0.3	4	0.2
0.4 - 0.6	19	0.5
0.7 - 0.9	13	0.8
1.0 - 1.2	4	1.1
1.3 - 1.5	0	1.4
1.6 - 1.8	0	1.7
	Total=40	

2.3 Cálculo de la media

La media se determina desde la fórmula:

$$\bar{X} = \frac{\sum f_i x_i}{N}$$

Incluyendo una columa adicional a la derecha de la tabla de frecuencia permite obtener en forma ordenada los productos $f.x$.

Ingresos (billones \$)	Frecuencia absoluta (f)	Marca de clase (x)	$f.x$
0.1 - 0.3	4	0.2	0.8
0.4 - 0.6	19	0.5	9.5
0.7 - 0.9	13	0.8	10.4
1.0 - 1.2	4	1.1	4.4
1.3 - 1.5	0	1.4	0
1.6 - 1.8	0	1.7	0
			Total=25.1

La suma de los productos de la última columa corresponde a calcular el término $\sum f_i x_i = 25.1$ y como hay $N = 40$ ingresos, la media es:

$$\bar{X} = \frac{25.1}{40} = 0.63$$

La media de los ingresos recibidos por empresa es de 0.63 billones de pesos.

2.4 Cálculo de la desviación estándar

La desviación estándar se determina desde la fórmula:

$$\sigma = \sqrt{\frac{\sum f_i x_i^2}{N} - \bar{X}^2}$$

En una nueva tabla de trabajo, se agrega una columna para realizar el producto $f.x^2$ o equivalentemente $f.x.x$.

Ingresos (billones \$)	Frecuencia absoluta (f)	Marca de clase (x)	$f.x^2$
0.1 - 0.3	4	0.2	0.16
0.4 - 0.6	19	0.5	4.75
0.7 - 0.9	13	0.8	8.32
1.0 - 1.2	4	1.1	4.84
1.3 - 1.5	0	1.4	0
1.6 - 1.8	0	1.7	0
			Total=18.07

La suma de los productos de la última columa corresponde a calcular el término $\sum f_i x_i^2 = 18.07$ que al dividir con $N = 40$ permite cálcular el primer término de la formula; en la anterior subsección se determinó la media con $\bar{X} = 0.63$, resultado a sustituir en el segundo término. De este modo, la desviación es:

$$\begin{aligned}\sigma &= \sqrt{\frac{18.07}{40} - 0.63^2}\\ &= \sqrt{0.45 - 0.39}\\ &= \sqrt{0.055}\\ &= 0.23\end{aligned}$$

Esto es, los ingresos registrados por las empresas están dispersados en 0.23 billones de pesos respecto a la media, por lo cual, la mayoria de empresas no superaron una barrera de 0.63+0.23=0.86 billones de pesos en sus ingresos.

2.5 Representación gráfica

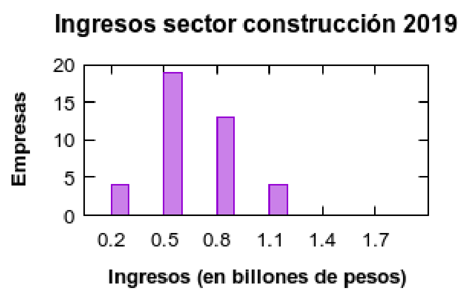

Figure 1: Histograma de los ingresos