Introducción al Análisis de Datos Agrupados

Breve resumen para estudiantes ITI

Mg. Carlos Andrés Pérez M.

Indice de contenidos

> **Advertencia**
>
> Este documento **NO ES DEFINITIVO** y se encuentra en permanente construcción. A medida que se vaya avanzando en la temática se irá complementando debidamente.

\newpage

1 Ideas generales

Al analizar datos estadísticos se precisa en ocasiones hacer *grupos*[1] para facilitar tanto el análisis de los datos como la presentación de resultados. A título de ejemplo no tiene mayor sentido ni resulta para nada práctico hacer un estudio sobre los ingresos económicos de los bogotanos y considerar aparte cada sueldo posible sino más bien hacer grupos predefinidos: menos de un millón, entre uno y dos millones, entre dos y tres millones, etc.

[1] A estos grupos se les conoce como **intervalos** o **clases** y, como es de esperarse, tienen una forma de abordarse acorde a su naturaleza.

Cabe decir que el **análisis de datos agrupados** (como se le llamará a este proceso de aquí en adelante) resulta necesario cuando se trabajan datos numéricos continuos y es a su vez la opción más adecuada para datos numéricos discretos cuando estos presentan una gran cantidad de valores distintos posibles.

IMPORTANTE: No olvidar que una variable es continua si corresponde a números que pueden tomar decimales y discreta si estrictamente toma valores enteros.

2 Tabla de frecuencias para datos no agrupados

> *A partir de aquí y para entender de la mejor manera cada parte del proceso, se ilustrará con ayuda de un ejemplo práctico:*

Supongamos que tenemos las estaturas (en metros) de 50 estudiantes y queremos analizar estadísticamente dichos datos:

```
 [1] 1.56 1.58 1.72 1.61 1.61 1.74 1.64 1.50 1.55 1.56 1.70 1.63 1.63 1.61 1.56
[16] 1.74 1.64 1.44 1.66 1.56 1.51 1.58 1.52 1.54 1.55 1.47 1.67 1.61 1.51 1.70
[31] 1.63 1.58 1.67 1.67 1.67 1.66 1.64 1.60 1.58 1.57 1.54 1.58 1.50 1.77 1.70
[46] 1.51 1.57 1.56 1.66 1.59
```

> **Nota**
>
> No debe preocuparse por los números de la izquierda (`[1]`, `[16]`, etc.) aquí o más adelante en este documento. Estos aparecen para facilitar en la lectura cuando se listan múltiples elementos. Por ejemplo, `[46]` nos ayuda a entender que el dato que sigue —es decir, `1.51`— es el que ocupa

la posición número 46 en la lista.

Ahora, por comodidad vamos a ordenarlos y observar qué nos vamos encontrando:

Aquí apreciamos fácilmente que mientras el estudiante más bajo mide 1.44 metros, el más alto 1.77 metros.

```
 [1] 1.44 1.47 1.50 1.50 1.51 1.51 1.51 1.52 1.54 1.54 1.55 1.55 1.56 1.56 1.56
[16] 1.56 1.56 1.57 1.57 1.58 1.58 1.58 1.58 1.58 1.59 1.60 1.61 1.61 1.61 1.61
[31] 1.63 1.63 1.63 1.64 1.64 1.64 1.66 1.66 1.66 1.67 1.67 1.67 1.67 1.70 1.70
[46] 1.70 1.72 1.74 1.74 1.77
```

Dado que resultaría muy poco práctico considerar cada valor de estatura entre estas dos medidas, haremos grupos.

Para ello primero necesitamos el **rango** (R) que no es otra cosa que la diferencia entre los valores máximo y mínimo de los datos: $1.77 - 1.44 = 0.33$. Esto se interpreta como que entre el estudiante más alto y el más bajo hay 0.33 metros de diferencia (33 cm).

Ahora, considerando este valor es que vamos a tomar la decisión más importante que nos acompañará durante todo el proceso: **cuántos intervalos o grupos debemos hacer** (cantidad que ahora llamaremos m)[2].

Para este ejercicio consideraremos 5 intervalos de igual amplitud c. Esta se calcula dividiendo el rango entre la cantidad de intervalos, es decir: $c = R/m$, que en nuestro caso sería $c = 0.33/5 = 0.066$ y la cual aproximaremos a 0.07 dado que ese es el nivel de precisión de nuestros datos (dos decimales).

> **Precaución**
>
> En estos casos SIEMPRE debemos redondear *por exceso* ya que si no estaríamos despreciando algunos de los datos originales.

Pero en este caso tendríamos un problema: al redondear la amplitud de los intervalos el rango original ya no nos serviría, por lo que nos tocaría ajustarlo. Veamos: El rango original era de 0.33 metros mientras que el nuevo sería de $m \cdot c = 5 \cdot 0.07 = 0.35$ metros. ¡Un descuadre de 0.02!

[2] La verdad es que no se trata de una decisión fácil al comienzo y es en última instancia escogido por quien analiza los datos a partir de su conocimiento de los mismos y experticia estadística. No obstante existen un par de reglas (en realidad más sugerencias que reglas) para apoyarnos: la *regla de la raíz* según la cual $m = \sqrt{N}$ y la *regla de Sturges* donde $m = 1 + \log_2 N$. Por ahora no las usaremos.

Esto no es problema ya que simplemente reajustamos los valores mínimo y máximo de manera equitativa y así se haría justicia: el nuevo mínimo sería 1.43 y el máximo 1.78 metros.

Ahora ya podemos ver que es posible ir de 1.43 a 1.78 en $m = 5$ saltos iguales de $c = 0.07$: $1.43, 1.50, 1.57, 1.64, 1.71, 1.78$.

Estos valores darán pie a la creación de nuestros intervalos: $[1.43, 1.50]$, $(1.50, 1.57]$, $(1.57, 1.64]$, $(1.64, 1.71]$, y $(1.71, 1.78]$ donde el paréntesis curvo "(" implica que el valor de la frontera no forma parte de dicho intervalo y el paréntesis cuadrado "]" indica que sí.

Ahora haremos un pequeño conteo de todos los valores para cada intervalo obteniendo la tabla que podemos observar a la derecha.

Tabla 1: Conteo para las estaturas en cada intervalo.

Intervalo	Frecuencia
[1.43,1.50]	4
(1.50,1.57]	15
(1.57,1.64]	17
(1.64,1.71]	10
(1.71,1.78]	4

La anterior tabla nos indica por ejemplo que de los 50 estudiantes hay 15 que miden de 1.51 a 1.57 metros, o que solo 4 miden de 1.72 a 1.78 metros.

Pero también resulta apropiado complementar la anterior tabla con los porcentajes respectivos. Esto se consigue dividiendo cada una de las frecuencias entre el total de datos (50 en este caso) y luego multiplicar por 100%. Esto a la final nos queda como:

Tabla 2: Tabla básica de distribución de frecuencias para datos agrupados.

Intervalo	Frecuencia	Porcentaje (%)
[1.43,1.50]	4	8
(1.50,1.57]	15	30
(1.57,1.64]	17	34
(1.64,1.71]	10	20
(1.71,1.78]	4	8

Y con esto ya tenemos una tabla de distribución de frecuencias súper básica para nuestros datos considerando cinco intervalos. Sin embargo, para ciertos propósitos como cálculos que haremos posteriormente (como los de las medidas de tendencia central)

esta tabla resulta poco práctica y conviene complementarla con más columnas así:

Tabla 3: Tabla de distribución de frecuencias para datos agrupados.

Intervalo	X_i	f_i	F_i	h_i (%)	H_i (%)
[1.43,1.50]	1.465	4	4	8	8
(1.50,1.57]	1.535	15	19	30	38
(1.57,1.64]	1.605	17	36	34	72
(1.64,1.71]	1.675	10	46	20	92
(1.71,1.78]	1.745	4	50	8	100

Donde X_i representa a la **marca de clase** el cual no es más que el valor medio entre los dos extremos de cada intervalo (es básicamente un valor que representa a todos los del intervalo); f_i y F_i las frecuencias absoluta y absoluta acumulada, respectivamente; h_i y H_i las frecuencias relativa y relativa acumulada expresadas como porcentaje, respectivamente.

3 Representación gráfica para los datos agrupados: Histogramas

Para representar datos agrupados la mejor forma es usando **histogramas**. Un histograma es, en esencia, un típico diagrama de barras con la particularidad de que cada barra se pega a la siguiente en el valor que sirve de frontera. Esto se hace para ilustrar la continuidad de los intervalos. La altura de la barra del histograma no es otra que la frecuencia absoluta respectiva.

En este orden de ideas lo que tendríamos es una gráfica como la de la derecha.

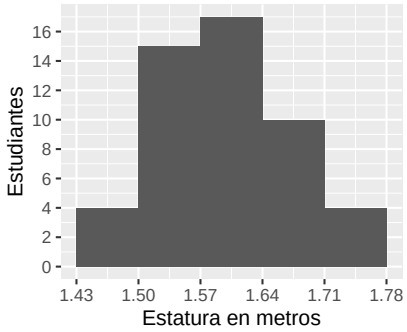

Figura 1: Histograma de la distribución de las estaturas.

4 Medidas de tendencia central para datos agrupados

En realidad las medidas de tendencia central (moda, mediana, media) se calculan exactamente de la misma manera que para datos no agrupados; sin embargo, en ocasiones ocurre que no tenemos los datos originales sino la tabla de datos agrupados ya hecha y a partir de ella es necesario establecer unos valores estimados que hagan de moda, mediana y media para estos datos para posteriores análisis.

En este orden de ideas consideraremos a partir de ahora el supuesto de que contamos no con los datos sino con la tabla siguiente y que ya antes hemos visto:

Intervalo	X_i	f_i	F_i	h_i (%)	H_i (%)
[1.43,1.50]	1.465	4	4	8	8
(1.50,1.57]	1.535	15	19	30	38
(1.57,1.64]	1.605	17	36	34	72
(1.64,1.71]	1.675	10	46	20	92
(1.71,1.78]	1.745	4	50	8	100

4.1 Moda

Dado que la moda es por definición el valor que más se repite dentro de un conjunto de datos, primero ubicaremos cuál es el intervalo que más se repite: en este caso $(1.57, 1.64]$ y lo llamaremos *intervalo modal*.

Para estimar la moda usaremos la siguiente fórmula:

$$M_o = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} c$$

Para entender esta fórmula consideremos lo siguiente: nuestro intervalo modal equivale al tercero, es decir, $i = 3$. En este orden de ideas $L_i = L_3 = 1.57$ (el límite inferior del intervalo), $f_i = f_3 = 17$, $f_{i-1} = f_2 = 15$ (la frecuencia absoluta del intervalo anterior), $f_{i+1} = f_4 = 10$ (la frecuencia absoluta del intervalo siguiente), y c por supuesto es la amplitud del intervalo ($c = 0.07$).

De esta manera tenemos que la moda estimada sería:

$$\begin{aligned} M_o &= 1.57 + \frac{17 - 15}{(17 - 15) + (17 - 10)} (0.07) \\ &= 1.57 + \frac{2}{2 + 7} (0.07) \\ &= 1.585556 \approx 1.59 \text{ metros} \end{aligned}$$

4.2 Mediana

Dado que la mediana corresponde al valor que divide a los datos en dos partes iguales, tenemos que mirar en la tabla en qué intervalo se alcanza el 50% de los datos y con esto tendríamos al **intervalo de la media**. En nuestro ejemplo sería el tercer intervalo ($i = 3$).

Para estimar la mediana para datos agrupados tendríamos que usar la siguiente fórmula:

$$M_e = L_i + \frac{N/2 - F_{i-1}}{f_i} c$$

N es la cantidad total de datos, F_{i-1} es la frecuencia absoluta acumulada para el intervalo anterior al intervalo de la media y los demás parámetros son los ya antes mencionados y trabajados.

Para nuestro problema tendríamos:

$$\begin{aligned} M_e &= 1.57 + \frac{50/2 - 19}{17}(0.07) \\ &= 1.57 + \frac{25 - 19}{17}(0.07) \\ &= 1.57 + \frac{6}{17}(0.07) \\ &= 1.594706 \approx 1.59 \text{ metros} \end{aligned}$$

4.3 Media

Para calcular la media a partir de los valores de una tabla de datos agrupados lo que debemos hacer es aplicar la siguiente fórmula:

$$\overline{x} = \sum \frac{X_i f_i}{N}$$

El símbolo $\sum$ corresponde a la letra griega *sigma* (en mayúscula) y en matemáticas se usa para indicar sumas.

En nuestro caso y para ilustrar la mecánica de la fórmula anterior tendríamos:

$$\begin{aligned} \overline{x} &= \frac{1.465 \cdot 4 + 1.535 \cdot 15 + 1.605 \cdot 17 + 1.675 \cdot 10 + 1.745 \cdot 4}{N} \\ &= \frac{5.860 + 23.025 + 27.285 + 16.750 + 6.980}{50} \\ &= \frac{79.9}{50} \\ &= 1.598 \text{ metros} \end{aligned}$$