Histograma

El histograma es una representación precisa de la distribución de un conjunto de datos numéricos. Es un gráfico de barras que se construye dividiendo el **rango** de los datos en intervalos (**bines**) iguales y contando cuantos datos caen dentro de cada intervalo. Sobre cada **bin** se dibuja un rectángulo cuya altura es proporcional a la cantidad de datos que caen dentro de cada **bin**. Los bines deben ser adyacentes y no deben superponerse.
El eje vertical puede normalizarse dividiendo la cantidad de datos en cada bin por el número total de datos. En ese caso, el histograma muestra la **frecuencia relativa** de los datos.

Ejemplo: se tiene el siguiente conjunto de datos correspondiente a la altura en metros de 30 estudiantes de un curso:

Estudiante	Altura [m]	Estudiante	Altura [m]
1	1.82	16	1.83
2	1.79	17	1.78
3	1.80	18	1.80
4	1.79	19	1.79
5	1.83	20	1.76
6	1.82	21	1.78
7	1.77	22	1.88
8	1.75	23	1.83
9	1.72	24	1.76
10	1.94	25	1.74
11	1.79	26	1.81
12	1.67	27	1.79
13	1.78	28	1.75
14	1.78	29	1.78
15	1.79	30	1.79

El rango de los datos es la diferencia entre la máxima y la mínima altura registrada. En este caso:

$$R = Alt_{Max} - Alt_{min} = 1{,}94\,m - 1{,}67\,m = 0{,}27\,m$$

No existe un criterio general para determinar el ancho óptimo o la cantidad de bines. Una forma usual de determinar el número de bines es utilizando la siguiente relación:

$$k = \sqrt{N}$$

donde k es el número de bines y N el número de datos. El número de bines debe ser un número entero, por lo que debe redondearse al entero más próximo. En este caso:

$$k = \sqrt{30} = 5{,}47 \approx 5$$

El ancho de cada bin se determina dividiendo el rango en la cantidad de bines:

$$h = \frac{R}{k} = \frac{0{,}27\,m}{5} = 0{,}054\,m$$

Por lo tanto, los límites de los bines para este ejemplo serán:

1,67	1,724	1,778	1,832	1,886	1,94

Finalmente hay que contar cuantos datos caen dentro de cada intervalo. Es decir, cuantos estudiantes tienen una altura comprendida entre 1,67m y 1,724m. Es importante no colocar un mismo dato en 2 bines consecutivos. Es decir, si hubiera un estudiante cuya altura es 1,778m podemos asignarlo al intervalo comprendido entre 1,724 y 1,778 o al intervalo 1,778 y 1,832, pero no a los 2 intervalos. No importa el criterio elegido, siempre y cuando se respete el mismo criterio para todos los intervalos.

El número de elementos en cada bin para el ejemplo anterior es:

2	6	20	1	1

El histograma finalmente queda como se muestra en la figura 1:

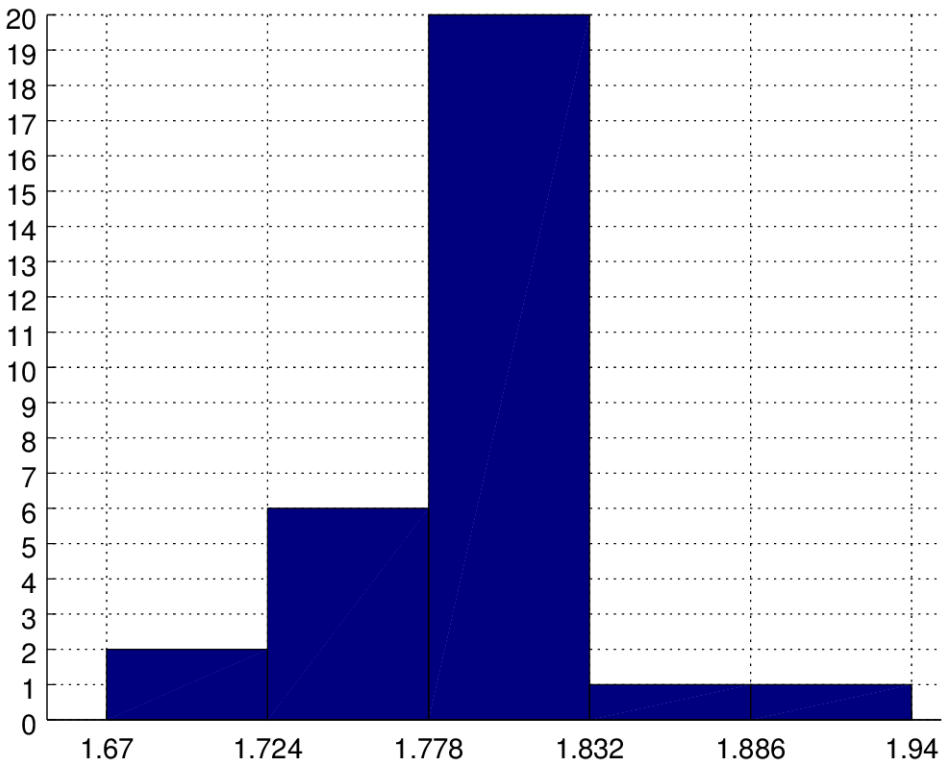


Figura 1: histograma.

Finalmente, si dividimos la cantidad de elementos en cada bin, por el total de datos, obtenemos las frecuencias relativas. La figura 2 muestra el **histograma normalizado**.

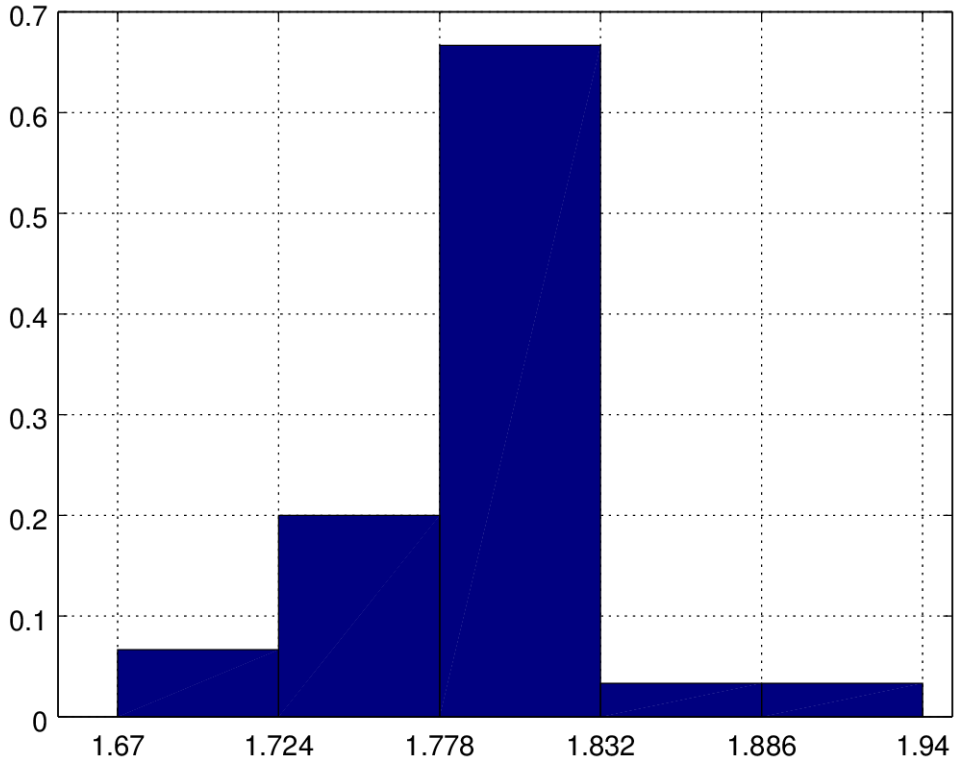
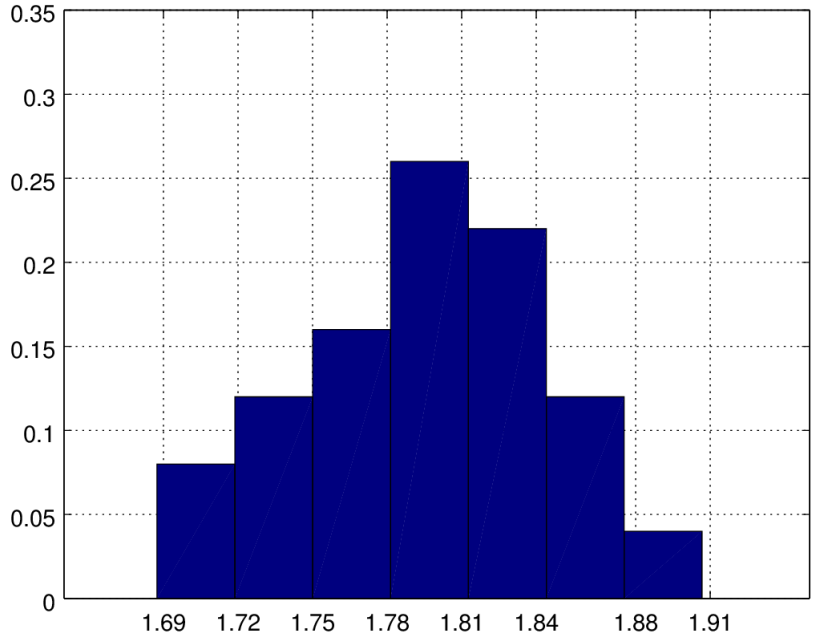


Figura 2: histograma normalizado.

A continuación se muestran los histogramas para conjunto de datos con 50 muestras, 200 muestras y 1000 muestras (figura 3). Se puede observar que a medida que aumenta el número de datos, el histograma se aproxima a una distribución Gaussiana.



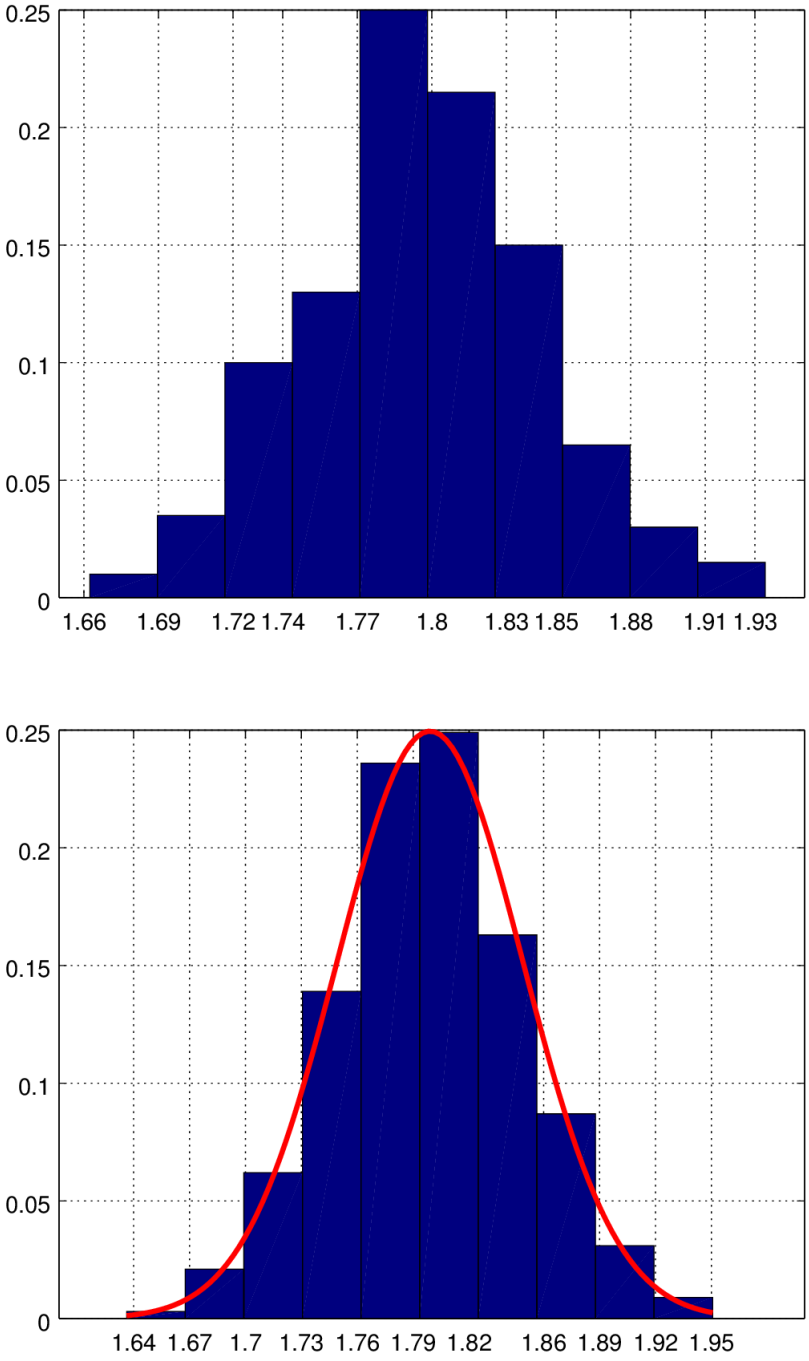


Figura 3: histograma para conjuntos de 50, 200 y 1000 datos con distribución normal de media igual a 1,80m y desviación estándar 0,05m.