PASOS PARA ARMAR LOS INTERVALOS DE UNA TABLA DE FRECUENCIAS

El siguiente es un enumerado de pasos y conceptos que seguramente le servirán como una ayuda para construir intervalos, tal como lo hicimos en el desarrollo del ejemplo anterior.

1) Localice los valores mínimo y máximo de la variable y calcule el rango

$$\boxed{R = x_M - x_m}$$

2) Tenga en cuenta que, cualquiera sea el número de intervalos que usted construirá, deberá **ampliar al menos mínimamente ese rango**, para garantizar que los valores mínimo y máximo "pertenezcan" a un intervalo (Si no lo hace, uno de los dos valores se "escapará" de los intervalos).

3) Decida cuántos intervalos va a construir (tenga en cuenta que no existe una receta mágica sino que debe usar su conocimiento del tema y su experiencia)

K = número de intervalos.

4) Comience a "probar" cuál será la amplitud de los intervalos, realizando el cociente entre el "rango mínimamente corregido" y el número de intervalos, hasta encontrar que ese **cociente de un número entero** (si es su objetivo) y se den algunas otras condiciones que a usted le interesen.

Las opciones son $\dfrac{x_M - (x_m - 1)}{K}$; $\dfrac{(x_M + 1) - x_m}{K}$;

o también $\dfrac{(x_M + 1) - (x_m - 1)}{K}$

Esto es, "corrija" uno ó los dos extremos, **ampliando** el rango y logre

$$\boxed{\frac{x_M^* - x_m^*}{K} = a}$$

Donde:

x_M^*: valor máximo corregido ($x_M^* \geq x_M$)
x_m^*: valor m'nimo corregido ($x_m^* \geq x_m$)
K : número de intervalos
a : amplitud de cada intervalo

Notas:

a. Observe que en el ejemplo se suma y/o resta "1". Esto se hace si los valores con los que está trabajando son enteros.
Puede también trabajar con décimos, centésimos, etc.

b. Algunos prefieren que el número "a" sea un número entero e **impar**, para garantizar que el número o valor que está en la mitad del intervalo, que se llama "centro o marca de clase", sea un número entero.

c. Algunos tratan, además, de que la marca de clase sea un número terminado en cero o cinco, para facilitar las operaciones aritméticas que luego se harán con ellos.

5) Tome el valor mínimo x^*_m (ya sea este un valor original o corregido) y construya el primer intervalo

$$\left.\begin{array}{r} x^*_m + a = l_s \\ \\ x^*_m = l_i \end{array}\right\} \qquad l_i \quad ; \quad l_s$$

6) Considere al límite superior del primer intervalo como límite inferior del segundo y repita el procedimiento hasta lograr el último intervalo, cuyo límite superior debe coincidir con

$$x^*_M$$

Al terminar de construir los intervalos de acuerdo con los pasos descriptos, seguramente ha advertido que el valor numérico del límite superior de cada intervalo (excluyendo el último) coincide con el valor numérico del límite inferior del intervalo siguiente.

¿Cómo procederá entonces si entre los valores observados de la variable, alguna coincide con esos valores numéricos?. ¿Dónde ubica ese valor de la variable?.

Puede resolver el problema de diferentes formas.

Si usted sólo modificó el valor mínimo pero no el máximo, esto es si trabaja con x^*_m y $x^*_M = x_M$, entonces deberá considerar cada intervalo semiabierto por la izquierda y hacer $\longrightarrow$ **(** l_i , l_s **]** para cada intervalo.

Ej.: **(** 28 , 33 **]** lo que implica que se incluya en ese intervalo el valor 33 pero no el 28.

Si por el contrario trabaja con x_m y x^*_M , entonces considerará $\longrightarrow$ **[** l_i , l_s **)** para cada intervalo.

Ej.: **[** 28 , 33 **)** lo que implica que se incluye el valor 28 pero no el 33.

Otra posibilidad es tomar el valor numérico de los límites y sumarles (o restarles) una mitad de la unidad de medida.

Ej.: **(**28,5 - 33,5**)** lo que implica que no se incluye el 28 pero si el 33, o bien (27,5 - 32,5) lo que implica que se incluye el 28 y no el 33.

Cualquiera sea la convención seguida para determinar límites de los intervalos, esta debe ser respetada a la larga de todo el análisis descriptivo de la información.

Los ítems mencionados significan (y en algunos casos seremos intencionalmente reiterativos) que

- tanto el x_m como el x_M **deben** pertenecer al primero y último intervalo respectivamente.
- todos los datos **deben** pertenecer a algún intervalo.
- los intervalos **no deben** tener valores comunes (deben ser mutuamente excluyentes).
- **de ser posible los intervalos se construirán de igual longitud**.
- es conveniente que la “marca de clase” o valor representativo del intervalo tenga la mayor cantidad de ceros posibles o esté terminado en cinco.
- el número de intervalos a determinar no será recibido como un “mensaje divino” “ni será” “adivinado graciosamente”; el número dependerá del problema del que se trate y del conocimiento que el “dueño de la información” tenga del tema. (Recordemos que en Estadística **no hay recetas mágicas**).
- es conveniente tratar de que no queden intervalos vacíos, si bien esto no siempre se puede lograr.

EJEMPLO PARA ANALIZAR (resuelto)

En una reunión de profesores jubilados de Estadística, Ana y sus amigos se encontraban comparando alegremente sus experiencias en los exámenes finales de Estadística de sus alumnos universitarios. Cada uno insistía en que sus alumnos eran los que "duraban menos" en los exámenes.

Antes de que la discusión se fuera a las manos, Ana propuso llevar a cabo una estadística macabra: para eso necesitaban tabular los resultados en forma adecuada (resultados obtenidos también en forma adecuada). He aquí los datos del "tiempo necesario para desaprobar a un alumno", eufemismo utilizado por la malvada Ana por "tiempo de duración de los alumnos" (expresión que "no suena" bien, ¿no es cierto?), en minutos:

27	24	25	17	3
43	9	34	48	33
28	28	0	40	27
19	27	15	38	32
45	19	35	32	0
33	21	35	16	13
16	39	13	19	35
43	44	11	11	31

x_m = 0 [min] (éstos duraron "lo que un respiro")
x_M^* = 48 [min] (fue el que opuso tenaz resistencia, aunque ya tenía su destino escrito)

$$\text{Rango} = R = x_M - x_m = 48 - 0 = 48min$$

Corrijamos este rango agregando una (1) unidad, esto es, un minuto más, al valor de x_M :

$$\text{Rango corregido} = R' = (x_M + 1) - x_m = (48 + 1) - 0 = 49min$$

Probemos con 10 intervalos: $k = 10$.

Entonces

$$a = \frac{R'}{K} = \frac{49}{K} = \frac{49}{10} = 4{,}9min$$

No nos conviene demasiado para hacer cálculos →entonces probemos con $K = 7$

$$a = \frac{R'}{K} = \frac{49}{K} = \frac{49}{7} = 7min$$

Esto significa que podremos trabajar con 7 intervalos de amplitud 7[min.].

Hallemos los límites inferiores y superiores de esos intervalos:

1º intervalo: $l_i = 0$ [*min.*]

$l_i + 7$ [*min.*]

2º intervalo: $l_i = 7$ [*min.*]

$l_s = 7 + 7$[*min.*]$=14$ [*min.*]

3º intervalo: $l_i = 14$ [*min.*]

$l_s = 21$ [*min.*]

4º intervalo: $l_i = 21$ [*min.*]

$l_s = 28$ [*min.*]

5º intervalo: $l_i = 28$ [*min.*]

$l_s = 35$ [*min.*]

6º intervalo: $l_i = 35$ [*min.*]

$l_s = 42$ [*min.*]

7º intervalo: $l_i = 42$ [*min.*]

$l_s = 49$ [*min.*]

Decidamos ahora qué sucederá con los valores que coincidan con los límites de los intervalos. Como el l_s del 7º intervalo es 49 [min.], está claro que no será un valor observando de la variable, porque $x_M = 48$ [min.]. Luego decidimos que los intervalos serán de la forma

$$[l_i \;;\; l_s)$$

Esto significa que el valor “conflictivo” se colocará en el intervalo en que aparezca como límite inferior.

Encolumnemos nuestros intervalos y hallemos su marca de clase o punto medio:

0 - 7	3,5
7 - 14	10,5
14 - 21	17,5
21 - 28	24,5
28 - 35	31,5
35 - 42	38,5
42 - 49	45,5

Esto podría significar un problema, si nos resulta difícil trabajar con decimales (por ejemplo, si no tenemos una calculadora a mano). Si es así, podemos cambiar la amplitud o el número de intervalos, o ambos.

Si ésta es su situación, le sugerimos que corrija nuevamente el rango ya corregido, sumando una nueva unidad; obtendrá así un $R' = 50$ min. , y podrá armar 5 intervalos de amplitud 10 min. o bien 10 intervalos de amplitud 5.
¿Con cuál de las opciones solucionará su problema?

Sigamos nosotros con el desarrollo anterior.

Contemos ahora las observaciones que caen en cada intervalo; usemos para ello palotes (o cruces, puntos, asteriscos, ...)

Intervalos	**M**arca de clase	**N**úmero de observaciones
0 - 7	3,5	III
7 - 14	10,5	~~IIII~~ I
14 - 21	17,5	~~IIII~~ III
21 - 28	24,5	~~IIII~~ II
28 - 35	31,5	~~IIII~~ III
35 - 42	38,5	~~IIII~~ II
42 - 49	45,5	~~IIII~~ I

Finalmente, podemos pasar en limpio los resultados obtenidos, con un poco más de formalismo

Límites de Intervalos (min.)	**M**arca de clase (min.)	**F**recuencia
0 - 7	3,5	3
7 - 14	10,5	5
14 - 21	17,5	7
21 - 28	24,5	6
28 - 35	31,5	8
35 - 42	38,5	6
42 - 49	45,5	5
	Total n =	40

Si hubiésemos decidido tomar 5 intervalos de amplitud 10 min., habríamos obtenido:

Límites de Intervalos (min.)	**M**arca de clase (min.)	**F**recuencia
0 a menos de 10	5	4
10 a menos de 20	15	11
20 a menos de 30	25	8
30 a menos de 40	35	11
40 a menos de 50	45	6
	Total n **=**	40

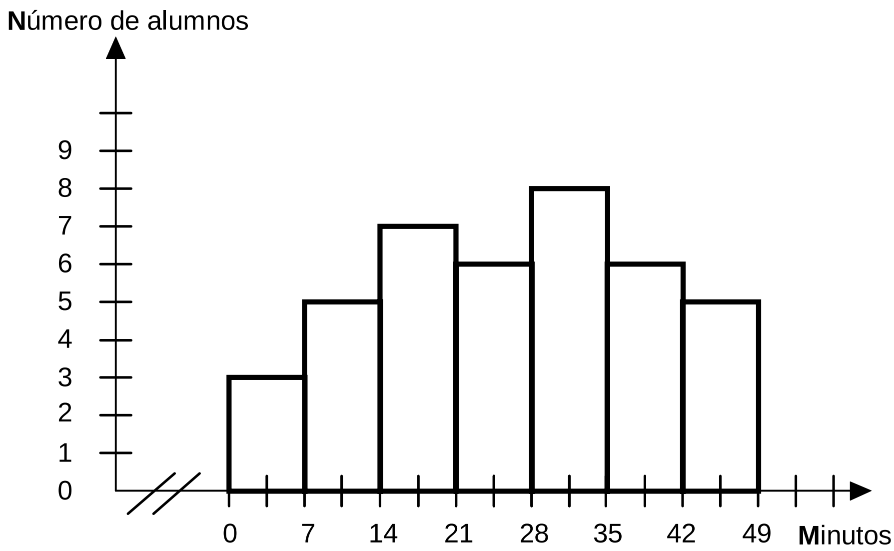
Observemos que al decir "0 a menos 10," 10 a menos 20", etc., estamos usando una expresión equivalente de [0, 10) ; [10, 20) ; etc., y mucho más adecuada para la representación ante personas que no manejen el idioma matemático. Además, nos estamos evitando el trabajo de aclarar con una nota al pie de la tabla la convención usada en el caso de las observaciones "problemáticas": en la primer tabla construida, deberíamos haber escrito al pie algo así como "los valores coincidentes con los límites de un intervalo se tomarán en el intervalo siguiente", o una expresión equivalente.

Observemos también que en las unidades agregadas para corregir el rango inicial sólo podían tomarse al final, y no al principio del conjunto de datos, puesto que no tendría sentido hablar de "alumnos que duraron - 1 min.". En otros problemas podrían tomarse al principio (x_m - 1 unidad , o x_m - 2 unidades), o uno al principio y otro al final (x_m -1 unidad y x_M + 1 unidad).

Finalmente, no debemos olvidar que estamos trabajando con datos reales, que tienen unidades: éstas deben colocarse siempre, para que las interpretaciones puedan hacerse correctamente sin necesidad de buscar más información que la dada por la tabla construida.

A pesar de que los intervalos que construimos tienen amplitud a = 7 min, pensemos por un instante que corresponden a 1 unidad (esto significa que: 7 min equivale a 1 unidad).

Si a cada intervalo (ahora "unitario") lo hacemos corresponder a la base de un rectángulo, y a la frecuencia la hacemos corresponder a la altura del rectángulo, vemos entonces que el área de ese rectángulo considerado tiene el mismo valor numérico que la frecuencia. Con esta transformación de frecuencias en áreas de intervalos de amplitud unitaria, podemos construir el histograma para esos datos:



Las dos líneas paralelas sobre el eje horizontal se colocan para indicar un "corte" en la escala que permita dibujar el histograma adecuadamente (sino habría quedado "pegado" al eje vertical). Si calculáramos el área del histograma, tendríamos:
que es precisamente el número de observaciones con las que contaban Ana y sus colegas al comienzo de este estudio estadístico.

$$\text{Area} = \sum_{i=1}^{7} (\text{base.altura})_i = (1.3) + (1.5) + (1.7) + (1.6) + (1.8) + (1.6) + (1.5) = 40$$

ALGUNAS INDICACIONES ACERCA DE CÓMO ELABORAR UN HISTOGRAMA DE FRECUENCIAS

Como ya se ha visto, el histograma no es más que una representación gráfica de una distribución de frecuencias de una variable cuantitativa continua, consistente en un conjunto de "rectángulos pegados".

El gráfico debe construirse teniendo en claro que **el área de cada rectángulo (y no su altura) representa a la frecuencia absoluta o a la proporción de elementos que pertenecen al intervalo**

El primer paso consiste en elaborar la escala vertical ($\vec{x}$) correspondiente a la variable en estudio.

Si observamos la tabla de frecuencias, los intervalos que hemos construido y por supuesto, tenemos en cuenta los valores mínimo (x_m) y máximo (x_M) de la variable, podemos armarla muy rápidamente, siempre teniendo en cuenta que la escala es numérica y que en consecuencia las distancias iguales deben representar siempre conjuntos iguales de valores.

Estamos suponiendo ahora que según "los pasos para armar los intervalos de una tabla de frecuencias" estos ya están listos y supongamos por el momento que son iguales (se siguieron los pasos del punto anterior).

Debemos ahora construir los rectángulos. La primera intención será, seguramente, igualar en forma directa la frecuencia absoluta o proporción (o porcentaje) correspondiente con la altura del rectángulo.

Es posible que así lo consideremos y que el hecho no sea un error, pero sólo será así si la base de cada uno de los rectángulos, **aún cuando coincida con un conjunto determinado de valores de la variable, sea considerada la unidad**; en este caso, la frecuencia no es realmente la altura, aún cuando la leamos en la escala vertical, sino que la frecuencia es el área del rectángulo.

Recordemos que en el histograma representamos las frecuencias absolutas o porcentajes como si estos se distribuyeran en forma homogénea a lo largo del intervalo considerado. Recordemos también que ésta es la razón por la cual el valor central, que llamamos "centro" o "marca de clase", es el valor de la variable que representa a todos los valores del intervalo.

¿CÓMO PROCEDEMOS SI LAS AMPLITUDES DE LOS INTERVALOS SON DIFERENTES?

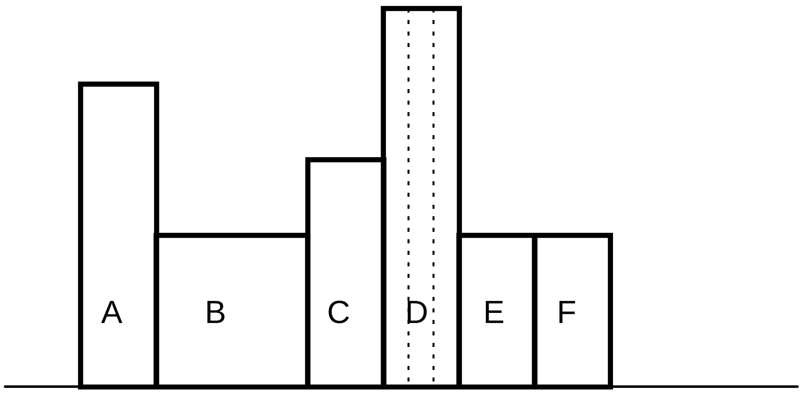
Para construir los intervalos en este caso y luego los rectángulos del histograma, es necesario reiterar que hay que tener en cuenta que la frecuencia del intervalo es el área del rectángulo correspondiente, y no la altura.

Se debe trabajar de modo que el área del rectángulo sea proporcional a la frecuencia del intervalo de clase.

RECUERDE . . . OBSERVE . . . RAZONE . . .
Importa el ÁREA y no la ALTURA . . .

Recordemos también que la diferencia entre un diagrama de barras (separadas) y un histograma (pegadas) es que en el primero, (que se emplea para variable cualitativa o cuantitativa pero en los casos en que los valores de la variable son considerados como “carteles” o rótulos que se diferencian no por su valor sino por su “tipo”) la frecuencia es directamente la altura de las barras. En un histograma, en cambio, por tratarse de variable cuantitativa continua, los rectángulos deben estar pegados y se lee la altura pero de acuerdo con la base del rectángulo considerado.

Si se tiene, por ejemplo



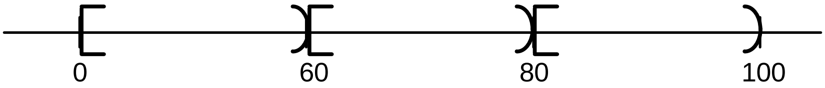
Se puede ver que la cantidad de elementos de A es prácticamente igual a la cantidad de elementos de B, ya que el primer intervalo es el doble de alto pero el segundo tiene una base doble a la anterior. Las áreas son similares por lo que podemos admitir que hay aproximadamente la misma cantidad de elementos en ambos rectángulos.

En el caso D, se observa que cada una de las tres partes tiene 1/3 de los elementos de D, ya que la base de cada rectangulito es la tercera parte del total.

Para tener en claro cómo proceder para construir un histograma cuando los intervalos a considerar son distintos, veamos un ejemplo.

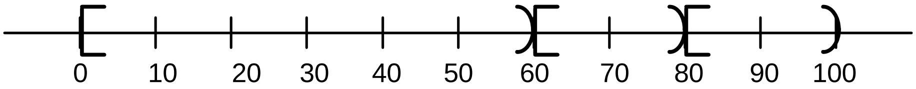
Supongamos que en la clase de Bioestadística, cursada 1997, 30 de los 60 alumnos inscriptos reprobaron, es decir que obtuvieron calificaciones menores a 60 puntos; 28 de los alumnos obtuvieron concepto, con calificación desde 60 a menos de 80 y el resto, los otros 12 alumnos, fueron promovidos sin necesidad de rendir el examen final. Además sabemos que ningún alumno obtuvo la calificación 100.

Si comenzamos a trabajar según lo visto anteriormente y “no pensamos mucho”, dibujaríamos una escala de puntajes como sigue

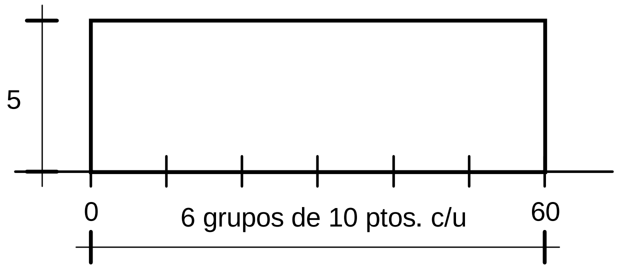


lo que supuesto no es correcto porque los segmentos dibujados son de igual longitud pero los puntajes incluidos no son iguales.

La escala correcta es:

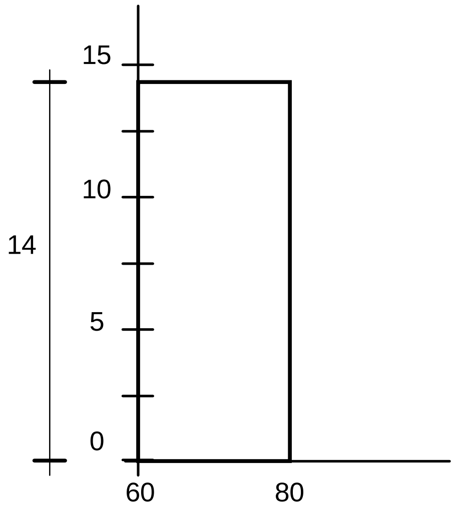


Razonemos ahora que 30 de los 60 alumnos obtuvieron calificaciones entre 0 y 60, pertenecientes a[0, 60), por lo que, si pensamos un poco y recordamos que los elementos se distribuyen en forma homogénea a la largo de cada intervalo, tendremos que si hay 30 elementos sobre 6 intervalos, tendremos 5 en cada uno de los intervalos, cada uno de ellos con 10 puntos de calificaciones como base, esto es

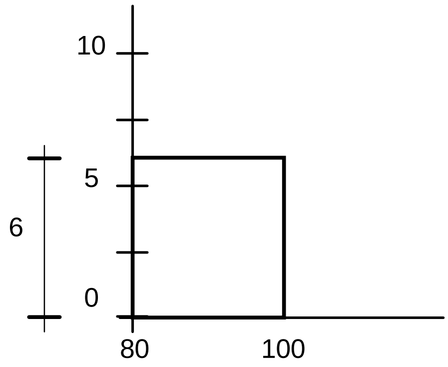


Nótese que se está considerando un conjunto de "10 puntos" de calificación como una unidad.

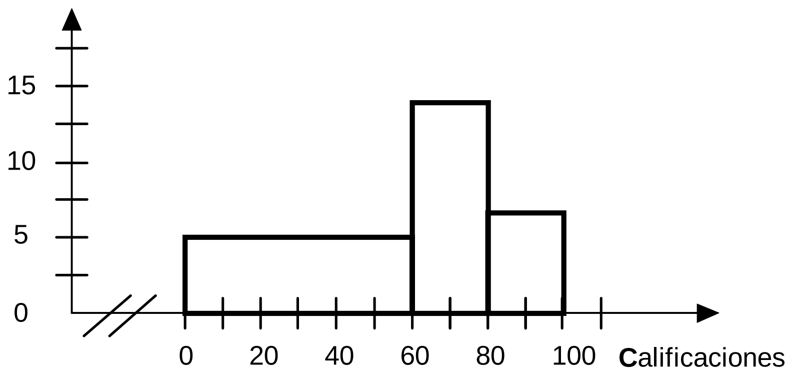
Se tiene además 28 alumnos con calificaciones entre 60 y 80 puntos, por lo que, con el mismo razonamiento, será:



y por último será, para los 12 alumnos con calificaciones de 80 a 100 puntos:



En definitiva, el histograma completo será:

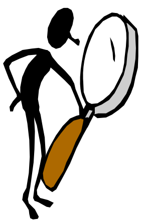


Cuando se interpreta un histograma de estas características es necesario tener mucho cuidado en “leer” la escala vertical. Esta escala puede presentar la frecuencia absoluta o bien el % de casos analizados.

Algunos autores opinan que:

Cuando se trabaja con una serie de frecuencias de variable continua marcadamente asimétrica, con unos pocos valores extremos, conviene construir intervalos cada vez de mayor amplitud cuando se van encontrando valores de la variable con escasa frecuencia, con la finalidad de evitar que aparezcan intervalos vacíos.

¿Usted que opina acerca de esta propuesta?



ATENCIÓN...
Lea los conceptos que sobre estos temas se emplean en LEVIN Y RUBIN - (citado en la bibliografía)
CRITIQUE - Razone y justifique sus críticas.