¿En cuantos intervalos conviene dividir los datos para construir un histograma? o, más en general, ¿qué aspectos hay que tener en cuenta para construir un histograma?

Roberto Behar Y Pere Grima.[1]

Abordaremos la respuesta a esta pregunta en tres etapas. En la primera nos referiremos a la situación más corriente en la que el todos los intervalos tienen el mismo ancho y observaremos como operan algunos programas como el Excel y el Minitab. Se harán algunas recomendaciones. Posteriormente y para esta misma situación de igual anchura de los intervalos, presentaremos los resultados de las reflexiones del profesor Ronald Fisher sobre este tema, que es enriquecido por una cuantificación de la eficiencia del agrupamiento, en relación con el ancho de los intervalos. Finalmente, se presentará la situación en la cual se considera la situación más general, en la cual se aceptan intervalos de diferente ancho, surgiendo así el importante concepto estadístico de densidad de frecuencia.

Es conveniente tener presente las motivaciones para agrupar lo datos en intervalos de clase. Las más importantes son:

- Claridad en la descripción.
- Facilidad de manipulación
- Lograr generar fórmulas que permitan realizar interpolaciones.

1. Histogramas con intervalos de ancho uniforme

Respecto al número de intervalos no hay una regla fija, aunque lo razonable es que su número aumente al ir aumentando el número de datos. Si se utiliza un programa de ordenador este ya dará un número de intervalos razonable. Si se hace a mano, una regla sencilla para tomar como referencia es la siguiente:

Núm. de datos	Núm. de intervalos
20* – 50	7
50 – 75	10
75 – 100	12
Más de 100	15

*Para menos de 20 datos es mejor utilizar un diagrama de puntos

[1] *Roberto Behar es profesor titular de la Escuela de Ingeniería Industrial y Estadística de la Universidad del Valle. Cali, Colombia*
Pere Grima, es profesor titular de la Escuela de Ingenieros Industriales de la Universidad Politécnica de Cataluña. Barcelona, España.

Pero también debe tenerse en cuenta que para facilitar la lectura del histograma es importante que la anchura de los intervalos sea un número sencillo. Por tanto, la tabla anterior se debe utilizar como primera aproximación, ya que el número exacto estará supeditado a tener un valor adecuado para la anchura de los intervalos.

Veamos a través de un ejemplo los aspectos más relevantes a tener en cuenta, tanto si los histogramas se construyen con ordenador como si se hacen a mano. La Tabla 1.1 contiene los pesos (en gramos) de 160 piezas de pan , 80 producidas con la máquina 1 y otras 80 con la máquina 2. El valor nominal es de 210 gramos, se considera tolerable una desviación de $\pm$ 10 y existe interés en conocer y comparar la variabilidad que presentan los pesos en ambas máquinas.

Tabla 1.1: Datos correspondientes al peso, en gramos, de 160 piezas de pan elaboradas con 2 máquinas

Máquina 1				Máquina 2			
209.2	209.5	210.2	212.0	214.3	221.8	214.6	214.4
208.5	208.7	206.2	207.8	215.3	216.7	212.3	212.0
204.2	210.2	210.5	205.9	215.7	213.8	215.2	202.7
204.0	203.3	198.2	199.9	212.5	210.2	211.3	210.4
209.6	203.7	213.2	209.6	208.4	214.9	212.8	214.8
208.1	207.9	211.0	206.2	212.3	216.2	208.4	210.8
205.2	204.8	198.7	205.8	208.1	211.9	212.9	209.0
199.0	197.7	202.0	213.1	207.5	209.9	210.6	212.3
197.2	210.6	199.5	215.3	206.9	207.1	213.6	212.2
199.1	207.2	200.8	201.2	209.6	209.5	206.8	214.2
204.6	207.0	200.8	204.6	212.2	209.8	207.6	212.6
214.7	207.5	205.8	200.9	211.4	211.2	214.4	212.6
204.1	196.6	204.6	199.4	209.6	209.2	206.1	207.1
200.2	205.5	208.0	202.7	203.5	206.9	210.6	212.3
201.1	209.2	205.5	200.0	209.1	206.3	209.8	211.4
201.3	203.1	196.3	205.5	208.0	207.9	205.3	203.6
202.2	204.4	202.1	206.6	210.0	209.4	209.1	207.0
194.1	211.0	208.4	202.6	215.6	211.8	205.4	209.0
204.8	201.3	208.4	212.3	214.5	207.5	212.9	204.3
200.6	202.3	204.3	201.4	209.1	205.8	212.0	204.2

La Figura 1.1 muestra los histogramas construidos con Excel (Excel 2000: Herramientas > Análisis de datos > Histograma) con todos los parámetros por defecto. En ambos casos aparecen 9 barras (deberían tocarse, ya que la variable representada es continua), pero lo más destacable es que los números que figuran en el eje horizontal son "raros" y esto dificulta su lectura y la interpretación del gráfico. Además, como todos los programas adaptan la escala al rango de variación de los datos, las escalas no son iguales, lo que complica la comparación.

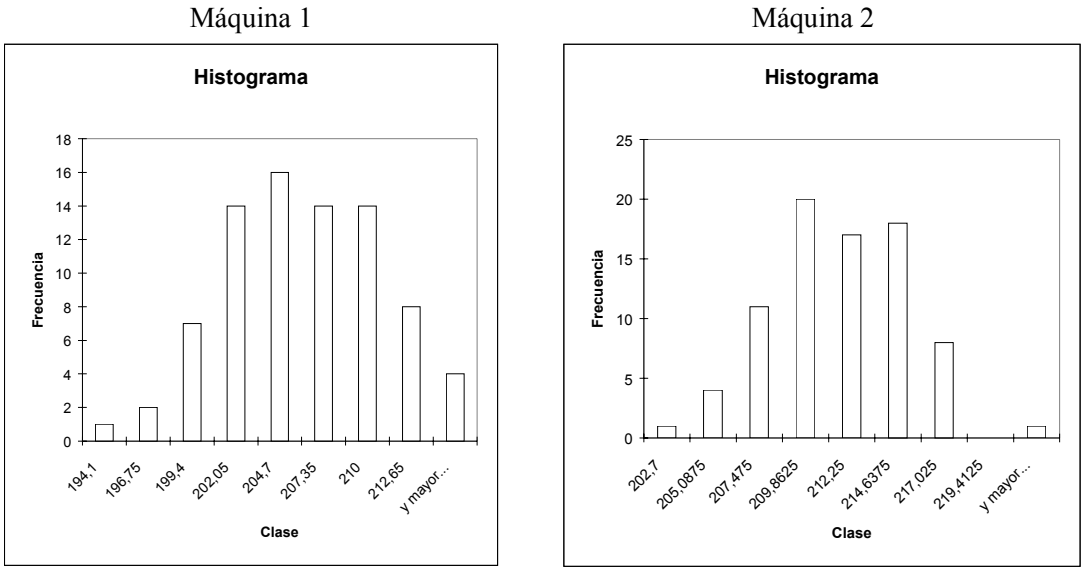


Figura 1.1: Histograma construido con Excel con todos los parámetros por defecto

Si utilizamos Minitab (Versión 13: Graph > Histogram) también con todos los parámetros por defecto, aparecen 12 barras para la máquina 1 y 11 para la 2. En este caso, tanto los números que figuran en los ejes como la anchura de los intervalos, son números fáciles y adecuados, aunque en el caso de la máquina 1 seguramente sería mejor tener más valores en el eje horizontal. En cuanto a las escalas, ocurre lo mismo que en el caso anterior.

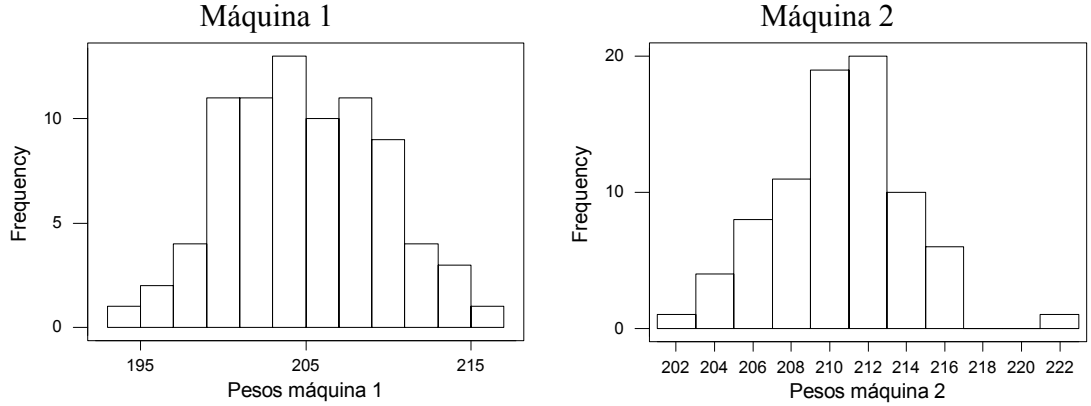


Figura 1.2: Histogramas construido con Minitab con los parámetros por defecto

Actuando sobre las opciones de Minitab se han construido los histogramas de la Figura 1.3, en los que los ejes están marcados con valores fáciles de leer, la anchura de los intervalos se ha mantenido en 2 gramos, y se ha forzado que las escalas sean iguales. También se han añadido unas líneas con el valor nominal y las tolerancias. De esta forma, sólo dando un vistazo se observa que la máquina 1 está descentrada mientras que la 2 está produciendo básicamente bien, aunque una unidad ha salido fuera de tolerancias por exceso.

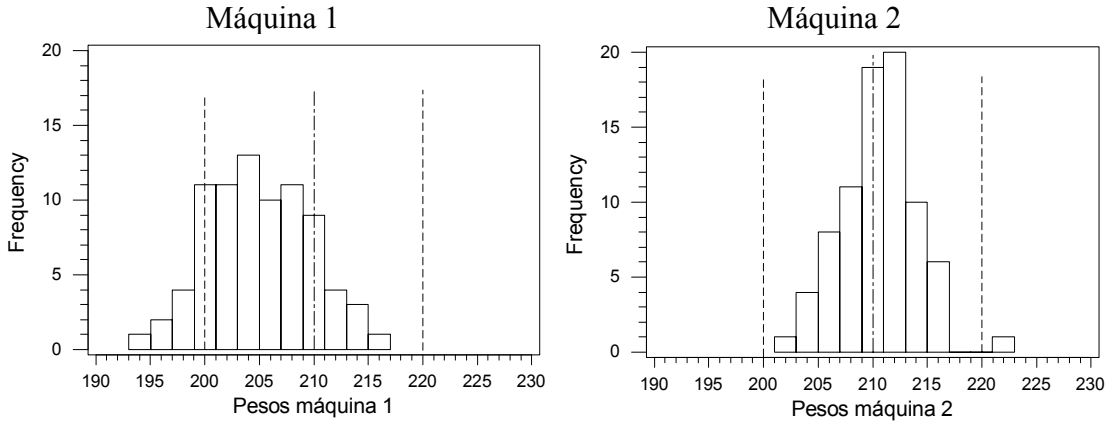


Figura 1.3: Histogramas construidos con Minitab actuando sobre las opciones disponibles para conseguir la apariencia deseada

En resumen, si el histograma se construye con un programa de ordenador, los aspectos a tener en cuenta para facilitar su lectura e interpretación son:

- Los ejes, especialmente el horizontal, deben estar marcados con valores fáciles de leer.
- La anchura de los intervalos también debe ser un número "redondo"
- Si se van a comparar varios histogramas, es necesario que todos ellos estén construidos con la misma escala para facilitar la comparación y evitar confusiones

Si estas características no aparecen con los parámetros que el programa tiene configurados por defecto, conviene actuar sobre las opciones disponibles para conseguirlo.

¿Y si se hace a mano?. En este caso los pasos a seguir son:

1. Calcular el rango de los datos. En el caso de la máquina 1, R = 215,3 – 194,1 = 21,2.
2. Plantear un número de intervalos en primera aproximación. En nuestro caso, con 80 datos, la tabla guía indica k =12 intervalos.
3. Calcular la anchura del intervalo, h, y ajustar a un número redondo. h = R/k, en nuestro caso h = 21,2/12 = 1,77, y por tanto lo más razonable es redondear a 2.
4. Tabular los datos de acuerdo con los intervalos definidos. Tener en cuenta que también interesa que los límites de los intervalos, o la marca de clase, sean números sencillos.
5. Construir el histograma. Si se va a comparar con otros, la escala debe ser lo suficiente amplia para que pueda ser común a todos ellos, y también conviene mantener en común tanto la anchura de los intervalos como sus extremos.

Una consideración para terminar. Seguramente estaremos de acuerdo en que construir histogramas a mano es una tarea un tanto tediosa, y si hacerlos con ordenador no es posible o implica gestiones y retrasos que no compensan, una buena idea puede ser utilizar una plantilla para recoger los datos de forma que al irlos anotando, el histograma se vaya construyendo sólo, tal como se indica, a título de ejemplo, en la Figura 1.4. Incluso aunque

se tenga ordenador disponible, la inmediatez en el análisis de los datos que se obtiene con este método puede hacer que sea el más adecuado.

Naturalmente, hay que tener una idea de por donde irá la variabilidad de los datos para poder diseñar la plantilla, que además, para su completa identificación y posibles análisis comparativos, siempre debe incluir un apartado con la fecha en que se han tomado los datos, su origen, la persona que los tomó, etc.

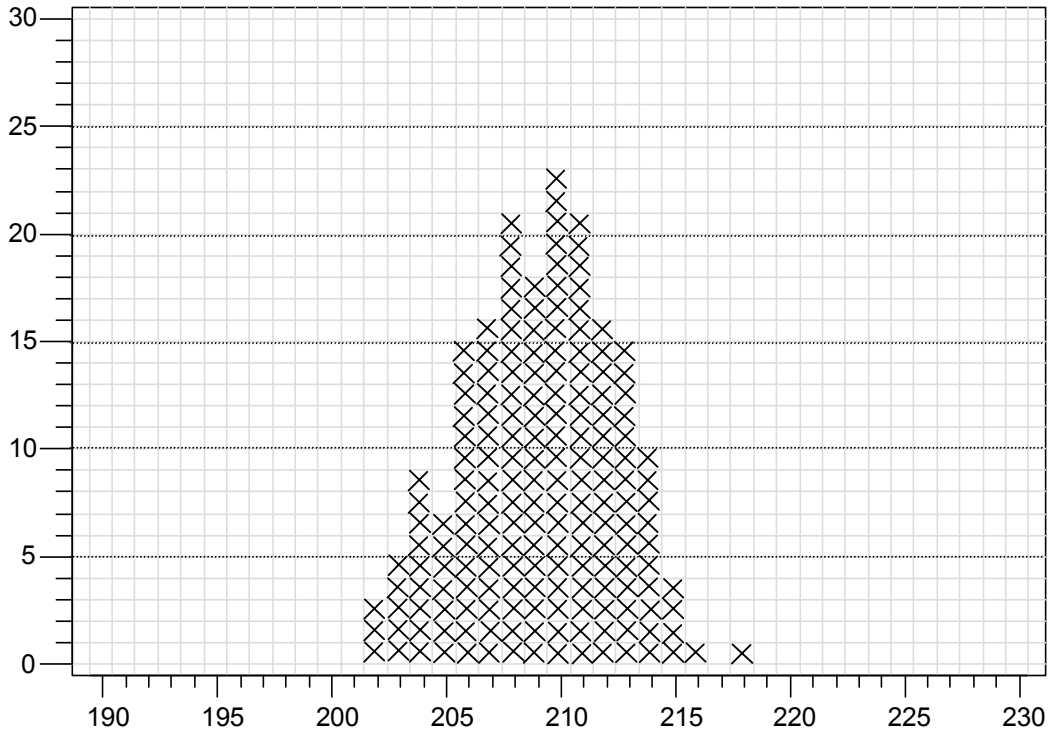


Figura 1.4: Plantilla de recogida de datos en la que el histograma se va construyendo solo. Los valores se redondean a las unidades y se marca una cruz en el lugar correspondiente.

2. Agrupamiento de los datos en intervalos de clase. El concepto de eficiencia de Fisher.

Cuando de antemano se sabe que los datos van a ser agrupados pueden elegirse instrumentos apropiados de medición que resulten más económicos, al no requerir una precisión exagerada, ni demasiadas cifras significativas.

Limitación del agrupamiento.

La mayoría de las herramientas matemáticas y estadísticas para el tratamiento de datos suponen que se dispone de observaciones individuales y supuestamente "exactas". Surge la pregunta: ¿En que medida pueden aplicarse estas herramientas a los datos agrupados?.

Intentando responder esta pregunta, puede decirse que en ocasiones pueden presentarse problemas de sesgo en las estimaciones. Teóricamente esto ocurre para cualquier tamaño de intervalo, sin

embargo cuando el ancho de los intervalos es menor que un cuarto de la desviación estándar, el sesgo es muy pequeño para propósitos prácticos.

¿Cómo influye el agrupamiento en la estimación?

Si nos referimos a una población con distribución normal y al estimador máximo verosímil de la media poblacional (que es la media aritmética), la eficiencia de la estimación con datos agrupados, con respecto a la que se obtiene con los datos originales depende del ancho de los intervalos.

Si el ancho de los intervalos fuese "hS" donde S es la desviación estándar y "h" un numero positivo, cuando el tamaño de muestra n es menos que 100 datos la eficiencia es como sigue:

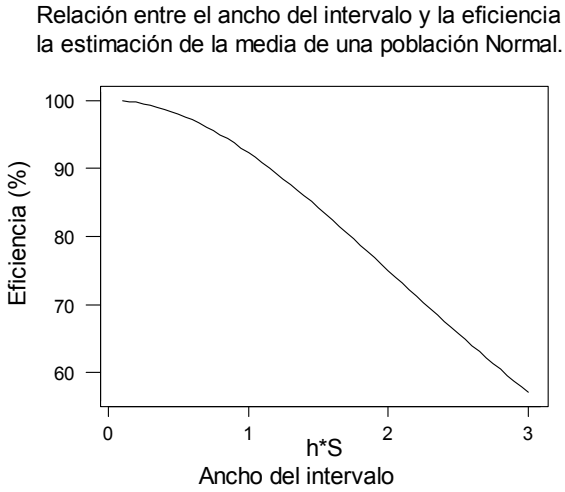
h	Eficiencia (%)
0.2	99.7
0.4	98.7
0.6	97.1
0.8	94.9
1.0	92.3

h	Eficiencia (%)
1.2	89.3
1.4	86.0
1.6	82.4
1.8	78.7
2.0	75.0

Estos valores se obtienen de la siguiente fórmula, dada por Fisher (1922).[2]

$$E = \frac{1}{1 + h^2/12}$$

La cual se muestra en el siguiente gráfico:



Para entender en significado de "Eficiencia", tomemos el caso de h=2, es decir intervalos de ancho 2 desviaciones estándar, el cual tiene asociada una eficiencia del 75%, que indica que para obtener

[2] Fisher Ronald (1922). " On Mathematical Foundations of Theorical Statistics. Pages 10308A-10368 In R. A. Fisher: "Contributions To Mathematical Statistics". New York: Wiley. First Published in the "Philosophical Transactions" Seie A. Volume 222 of the Royal Society of London.

la información sobre la media poblacional que se logra con 75 observaciones desagrupadas, se requerirían 100 agrupadas.

Por otro lado, el estimador[3] de la desviación estándar tiene una eficiencia del 58% cuando h=2.

Para estimar la desviación estándar poblacional σ los intervalos de un ancho 1.6σ, producen una eficiencia del 70%.

Conclusiones prácticas.

1. No invierta demasiada energía en afinar las mediciones para obtener datos que van a ser agrupadas, dicha energía adicional inviértala en más datos menos precisos.
2. Suponiendo que el rango en una distribución normal es 6 desviaciones estándar, entonces podemos asociar a cada valor "h", un número de intervalos "m", de la siguiente forma:

$$m = \frac{6\sigma}{h\sigma} = \frac{6}{h}$$

   Así por ejemplo h=0.2 es equivalente a toma 30 intervalos y h= 0.5 es equivalente a tomar 12 intervalos. En este último caso se tendría una eficiencia de 97.9%.

3. Corrección de Sheppard.

   El mejor estimador de la varianza σ^2 para datos agrupados es:

$$S^2\left(1-\frac{h^2}{12}\right)$$

Sin embargo cuando se calcula la media y la varianza con el mismo agrupamiento (intervalos de igual ancho), para pruebas de hipótesis no es necesaria la corrección de Sheppard, pues se presenta un efecto cancelativo.

Referencia adicional sobre el tema:

> Gjddbeh N. F. (1949). "*Contributions to the study of grouped observations: I. Application on the method of maximun likelihood in case of Normality distributed observations". Skandivish Actuarietidsrift 32. 135-159.*

3. Tienen que ser todos los intervalos necesariamente del mismo ancho?: El concepto de densidad de frecuencia.

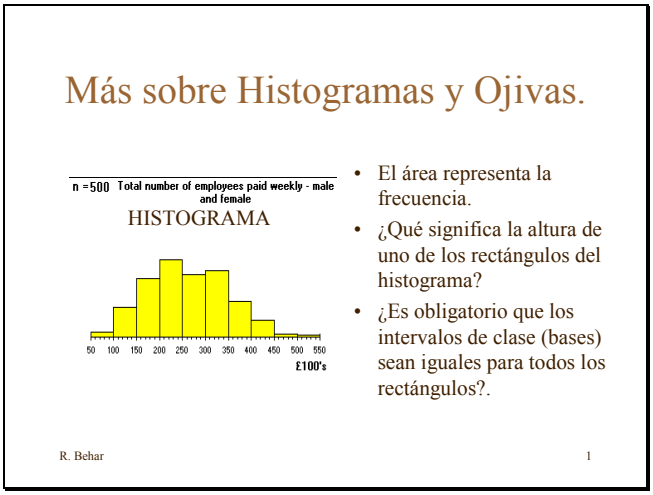
No, no siempre debe ser del mismo ancho, algunas situaciones obligan a usar intervalos de distinto ancho. Por ejemplo cuando se trata de variables como "salarios", a menudo es conveniente usar ancho distinto, pues para salarios bajos, pequeñas diferencias son importantes, pero las mismas no lo son para salarios altos. Así por ejemplo para un salario de 300 unidades monetarias, 30 unidades representa el 10%, sin embargo para salarios de

[3] Estimador de Máxima Verosimilitud.

6000 unidades, esta diferencia deja de ser importante. En este caso sería recomendable, usar intervalos cortos al principio de la escala e ir aumentando su tamaño.

En estos casos de intervalos de diferente tamaño, es necesario tener en cuenta que la escala del eje vertical del histograma, debe corresponder con la densidad de frecuencia y no con la frecuencia, pues lo que informa en un histograma es el área y no las alturas.

Cuando los intervalos son del mismo ancho las alturas son proporcionales al área y las escalas en el eje Y, al usar frecuencia o densidad, solo se diferencian por un factor constante, no produciendo distorsión en el grafico resultante.



El histograma es una de las representaciones gráficas de las distribuciones de frecuencias, más tradicionales y más usadas. Su relación conceptual y directa con conceptos como la función de densidad de probabilidad, justifica explorar un poco más aspectos que son ignorados en todos los textos de estadística que se disponen en el mercado.

Ya hemos dicho, que el histograma se interpreta bajo la premisa de que el área representa la frecuencia, no la altura de los rectángulos.

Cuando los intervalos de clase, son de igual anchura, el área es directamente proporcional a la altura y en este caso las alturas son un indicador de la frecuencia. Pero, esto se da en este caso particular y no vale cuando los intervalos son de diferente ancho, como puede suceder en muchas situaciones de carácter práctico.

Veamos un ejemplo:

Ejemplo: Antigüedad en el trabajo.

- En el sector de la industria metalmecánica, se toma una muestra al azar de 500 obreros y se determina la antigüedad en su trabajo.
- Por razones de índole administrativo, se quiere representar los datos por medio de un histograma que considere los siguientes intervalos de clase: 0-2 años, 2-3 años, 3-5 años, 5-10 años, 10-20 años.

R. Behar 2

Disponemos en principio de 500 datos sobre la antigüedad de los trabajadores que constituyen la muestra, cada uno de los cuales es clasificado en alguno de los intervalos propuestos. Luego de contar cuantos resultan en cada *intervalo de clase*, podemos construir un cuadro de frecuencias, como se muestra a continuación:

Cuadro de frecuencias para "antigüedad"

Antigüedad (Años) (I. Clase)	Frecuenc Absoluta	Frec. Relat.(%)		
0-2	50	10%		
2-3	25	5%		
3-5	200	40%		
5-10	200	40%		
10-20	25	5%		
TOTAL	**500**	100%		

R. 3

El número de datos en un intervalo es su *frecuencia absoluta* y si se expresa como un porcentaje del total de datos (500), se da origen a la llamada *frecuencia relativa.*

Este cuadro es la base para construir el histograma que los represente.

¿Cómo construir el histograma? En ese caso colocaremos en el eje X, la variable antigüedad de tal manera que dispongamos de una escala que cubra de o a 20 años.

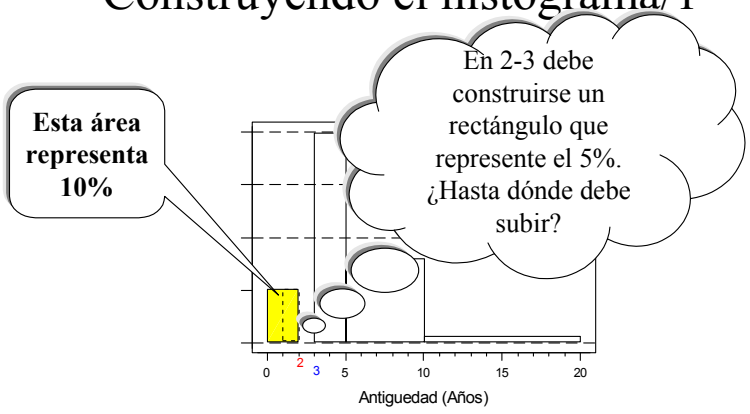
Bases para la construcción del histograma.

- Cada rectángulo se construye sobre su intervalo correspondiente.
- El área del rectángulo sobre un intervalo, representa su frecuencia. (% de datos).
- Por lo anterior: si un intervalo contiene el doble de datos que otro, debe estar representado por un rectángulo del doble de área.

R. Behar 4

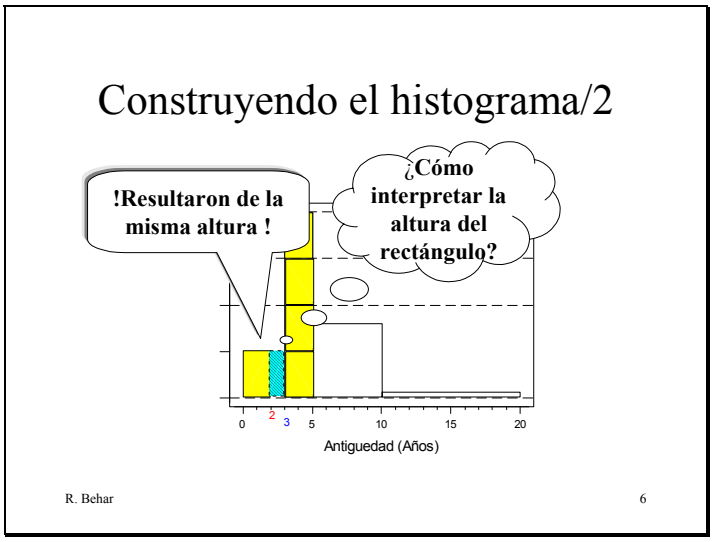
En nuestro caso podemos construir por ejemplo, el primer rectángulo con base en el primer intervalo 0-2 años y le colocaremos una altura arbitraria. El area que resulte es equivalente al 10% de los datos (su frecuencia relativa). Este servirá de patrón para construir los rectángulos restantes que constituyen el histograma.

Construyendo el histograma/1



R. Behar 5

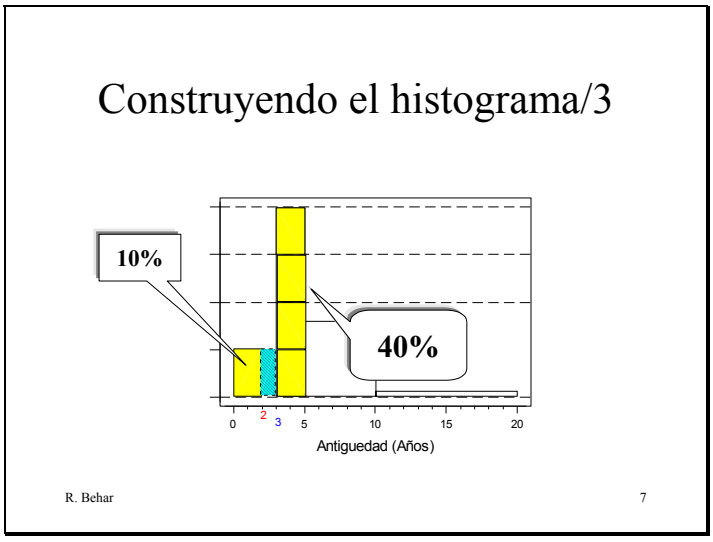
De esta manera, puesto que en el intervalo que sigue el que va de 2 a 3 años de antiguedad contiene 5% de los datos, el aéra de su rectángulo debe ser la mitad del anterior, para ello extenderemos su altura hasta donde se necesario para lograrlo.



Como en ancho del segundo rectángulo es la mita del primero, poniendo su misma altura tendremos la mitad del área, que representa el 5% de los datos.

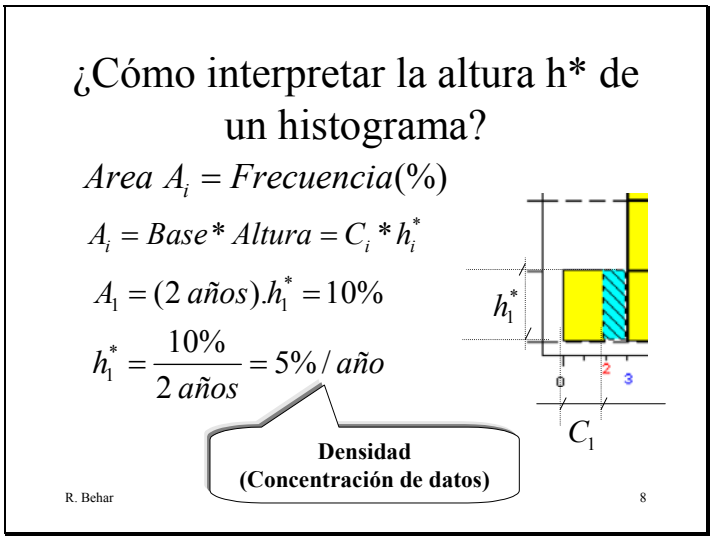
Note que no obstante que el primer intervalo contiene el doble de datos que el segundo, tienen, en este caso, la misma altura. Este hecho debe llevarnos a la reflexión sobre el significado de la altura de un histograma (eje Y).

El tercer intervalo, de 3 a 5 años, contiene el 40% de los datos, así, en el histograma, su rectángulo correspondiente debe ser representado por un área equivalente a 4 veces la del primero que tiene 10%, como se muestra a continuación:



En estos casos ha sido relativamente sencillo, casi a ojo, asignar la altura adecuada para obtener el área deseada, pero no siempre será tan obvia, la situación.

Intentemos sistematizar el cálculo de la altura de los distintos rectángulos y de paso ganemos en su interpretación, que nos permitirá realizar estimaciones de una manera muy sencilla y útil.



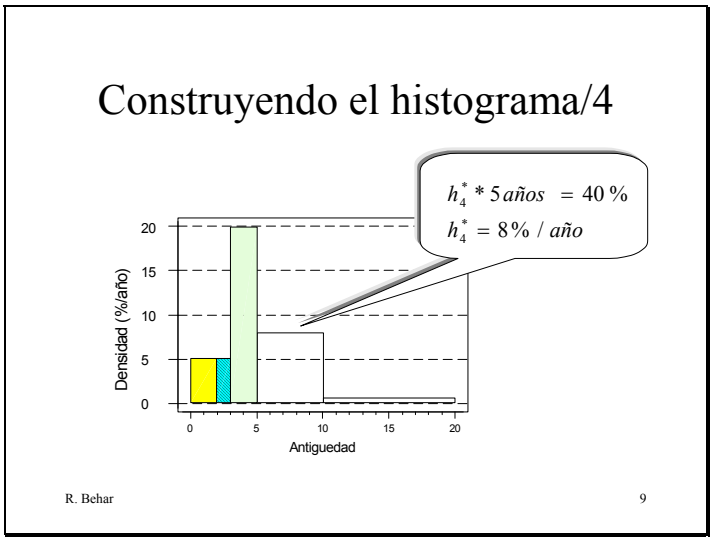
Si queremos que el área coincida con la frecuencia relativa, al *escribir la ecuación: base por altura igual a porcentaje de datos*, se puede despejar la altura que debe ponerse al rectángulo.

NOTACIÓN : La altura la representamos con h^* (h con asterisco), reservando la h (sin asterisco) para el porcentaje de datos que contiene el intervalo.

Así pues, para el primer intervalo: $h_1 = 10\%$ mientras que $h_1^* = 5\%/año$.

Practiquemos lo mencionado para calcular la altura h^*_4 del rectángulo asociado al cuarto intervalo.

Planteemos la ecuación *base por altura igual a porcentaje de datos* y despejemos la altura correspondiente:



Observemos que en el primer intervalo había 10% de los datos repartidos en dos unidades (dos años), lo cual asigna en promedio 5% de los datos a cada una de las dos unidades $h_1^* = 5\%/año$. El segundo intervalo contiene 5% de los datos, distribuidos en una unidad (un año), lo cual da $h_2^* = 5\%/año$.

El tercer intervalo (3 a 5 años), contiene el 20% de los datos, distribuidos en dos años, lo cual indica que en promedio en cada unidad (año), hay $h_3^* = 10\%/año$.

Para el cuarto intervalo, en el cual hay el 40% de los datos en 5 unidades (años), habrá en promedio $h_4^* = 8\%/año$ y en el último, 5% en 10 años, $h_5^* = 0.5\%/año$.

El siguiente cuadro presenta el resumen de la situación:

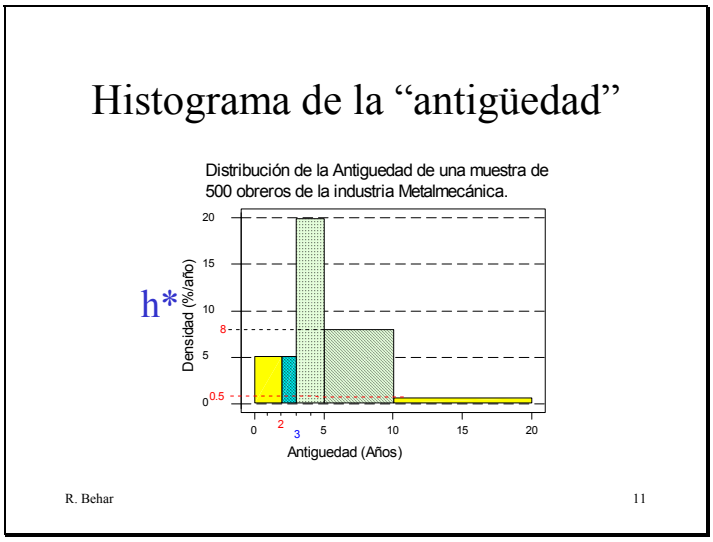
Cuadro de frecuencias para "antigüedad"

Antigüedad (Años)	Frecuenc Absoluta. n_i	Frec. Relat.(%) h_i	Densidad de frecuencia $h_i^* = h_i / C_i$
0-2	50	10%	$h_1^* = \frac{h_1}{C_1} = \frac{10\%}{2\ años} = 5\%/año$
2-3	25	5%	5%/año
3-5	200	40%	20%/año
5-10	200	40%	8%/año
10-20	25	5%	0.5%/año
TOTAL		100%	

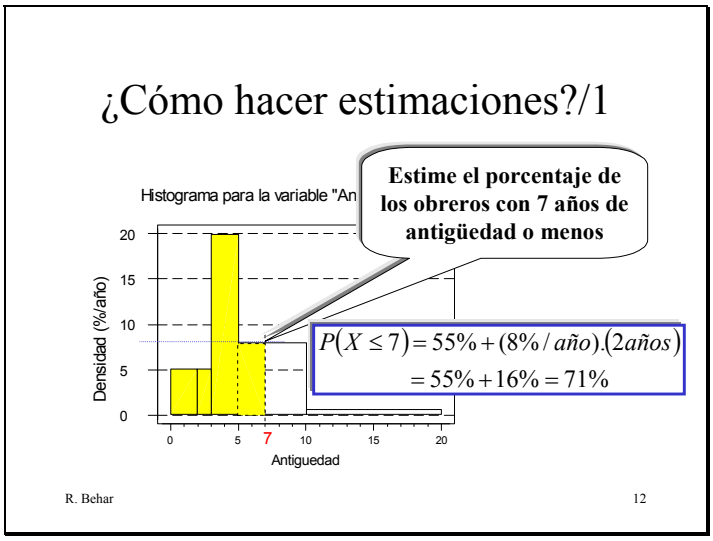
R. 10

En síntesis, la altura h_i^* de un rectángulo en un histograma, representa en promedio la *densidad (concentración) de los datos*, en dicho intervalo, expresado en %/unidad.

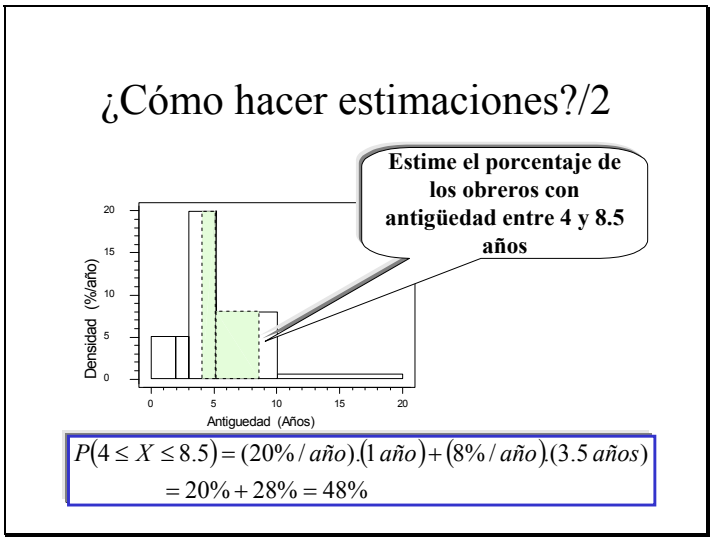
Sabiendo que la mayor densidad calculada es 20%/año, podemos metrizar el eje Y, de 0 a 20%/año y construir el histograma correspondiente.



Entender lo que significa una densidad de frecuencia permitirá que tenga sentido la función de densidad de probabilidad. El concepto de densidad, permite también hacer estimaciones muy razonables de cantidades no explícitas, en los datos agrupados o complicadas de realizar con base en una muestra bruta.

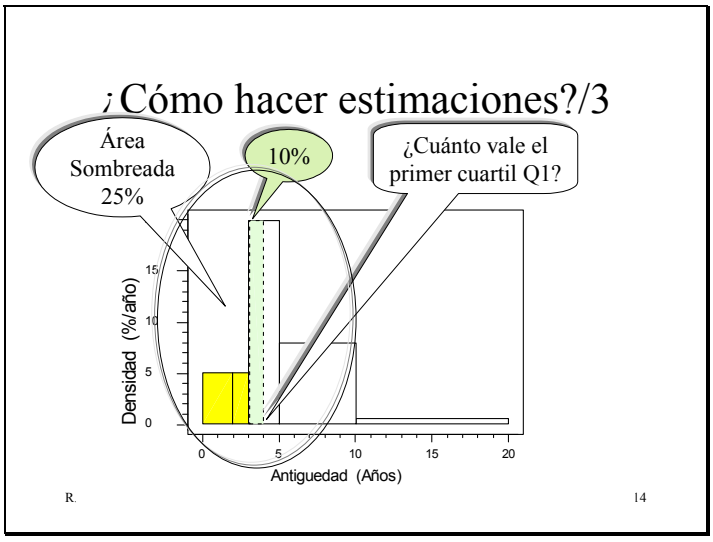


Si queremos estimar el porcentaje de trabajadores con antigüedad menor o igual que 7 años, hacemos la siguiente reflexión : "Con antigüedad de 5 años o menos hay 55% (es decir 10%+5%+40%), falta por sumar los que tienen antigüedad de 5 a 7 años, pero como en ese intervalo hay 8%/año, en los dos años que requerimos habrá 16%.
Así, el porcentaje pedido es 71% (Es decir, 55%+16%).



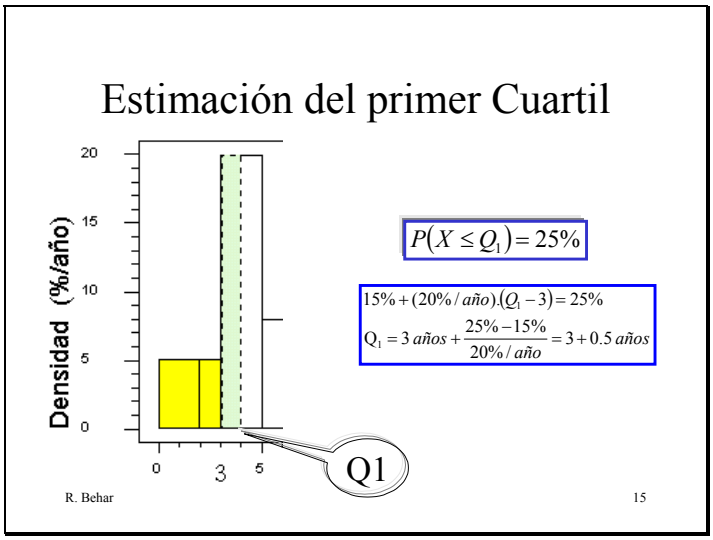
Este intervalo toma un año del tercer intervalo y 3.5 años del cuarto intervalo. Multiplicando este número de unidades por sus densidades respectivas y sumando, se da respuesta a lo pedido.

Hagamos un ejercicio más usando el útil concepto de densidad. Calculemos el primer cuartil Q1 para la antigüedad, es decir cual es el valor de la antigüedad tal que el 25% de los trabajadores no la superan.



En este caso conocemos el área (25%) y nos piden al valor Q1. Como se conocen los porcentajes de datos en todos los intervalos, podemos determinar con certeza, cual de los intervalos contiene el valor Q1 que estamos buscando. En este caso atrás de 3 años está el 15% de los datos y atrás de 5 años, está el 55% de los datos, por lo tanto en algún punto entre 3 y 5 deberá estar aquel que deja atrás el 25%.

Puesto que hasta 3 hay 15%, debemos saber cual es la distancia (Q1-3), que atrapa el 10% de los datos, para ser adicionada al valor 3 años, como se explica:



Lo que muestra la ecuación de arriba es que el 25% se debe completar sumando al 15% de los datos que están en los primeros dos intervalos, el porcentaje que se halla desde 3 hasta Q1. De allí puede despejarse Q1=3.5 años. Lo cual significa que la cuarta parte los trabajadores tienen 3.5 años de antigüedad o menos.
Para verificar que ha comprendido, calcule el segundo y tercer cuartiles y compruebe que ellos son respectivamente: 4.75 años y 7.5 años.

Observaciones sobre el histograma:

- El histograma es un modelo para el conjunto de datos. Como en todo modelo, se hacen simplificaciones con el propósito descubrir algunos rasgos asociados con el cuerpo de datos. En particular, se tiene una foto que describe el bosque, pero que impide ver cada árbol en particular (datos).
- En la construcción del histograma, hemos supuesto que la densidad (altura de los rectángulos) es constante para cada unidad dentro de un intervalo, es decir, consideramos que dentro de un intervalo, los datos se hallan esparcidos uniformemente, como cuando extendemos mantequilla sobre un pan. Esto es una de las simplificaciones, que nos permiten hacer cálculos y estimaciones muy fácilmente, esperando que el verdadero valor que resultaría, si tomáramos en cuenta los datos crudos originales, no sea muy distinto. Esto será mas cierto entre mayor sea el número de intervalos, pero hará cada vez menos sencillo el modelo.
- La mismas simplificaciones incluidas en el modelo de histograma, permite hacer más sencilla la estimación de rasgos de la distribución como la media, los cuartiles o la desviación estándar, entre otras.