Resumiendo datos continuos: el histograma

Ejemplo 22 *Los datos son el número de espectadores en 32 partidos del equipo nacional (en miles).*

42,1	51,0	30,0	35,2	29,3	10,9	16,1	51,6
47,0	51,4	35,2	31,7	17,8	67,0	43,2	23,7
25,2	36,1	32,3	51,7	46,0	12,2	21,1	29,0
14,3	47,2	31,3	35,4	29,1	23,0	10,3	34,2

En primer lugar, dividimos los datos en intervalos de igual anchura.

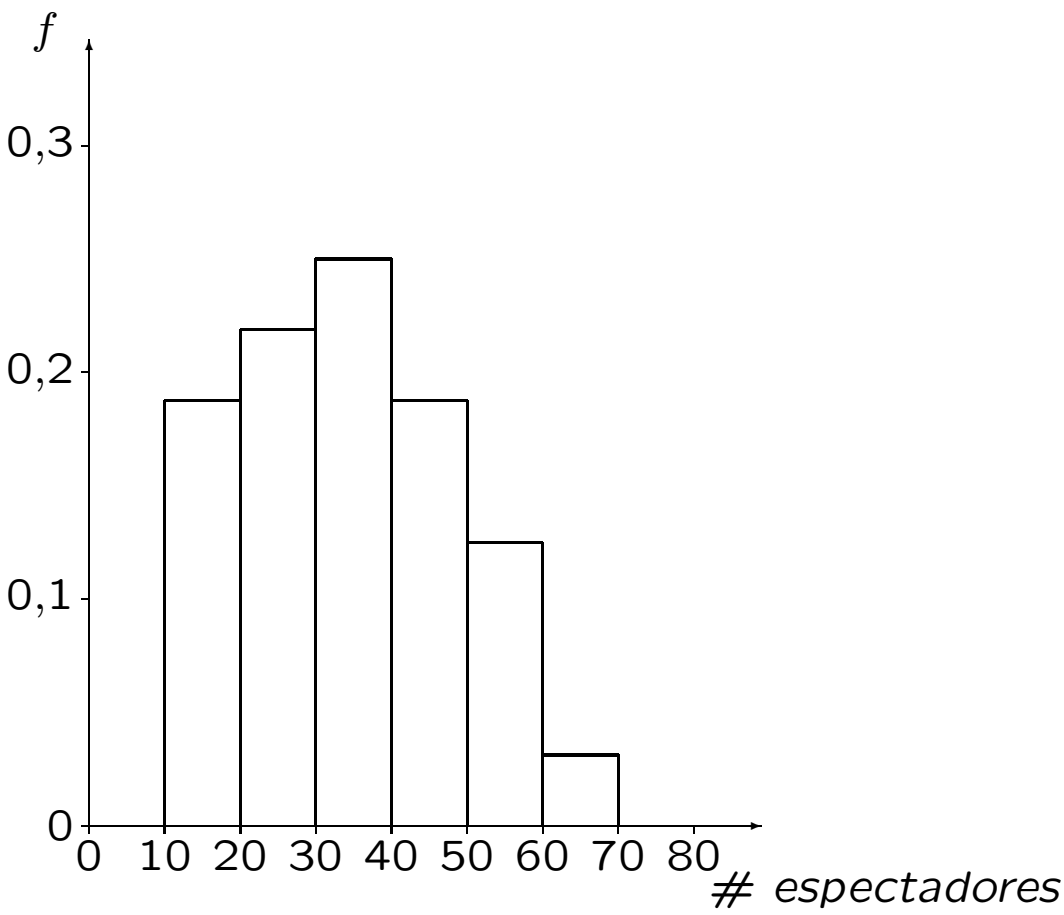
El valor mínimo de la muestra es 10,3 *y el valor máximo es* 67,0. *Ponemos intervalos de anchura* 10 *y construimos una tabla de frecuencias.*

*Tenemos mucho guidado en clasificar los valores en el límite de los intervalos (*30,0*)*

Clase	n_i	f_i
$[10, 20)$	6	0,1875
$[20, 30)$	7	0,21875
$[30, 40)$	8	0,25
$[40, 50)$	6	0,1875
$[50, 60)$	4	0,125
$[60, 70)$	1	0,03125
> 70	0	0
Total	32	1

Ahora, construimos el histograma.

Histograma de números de espectadores

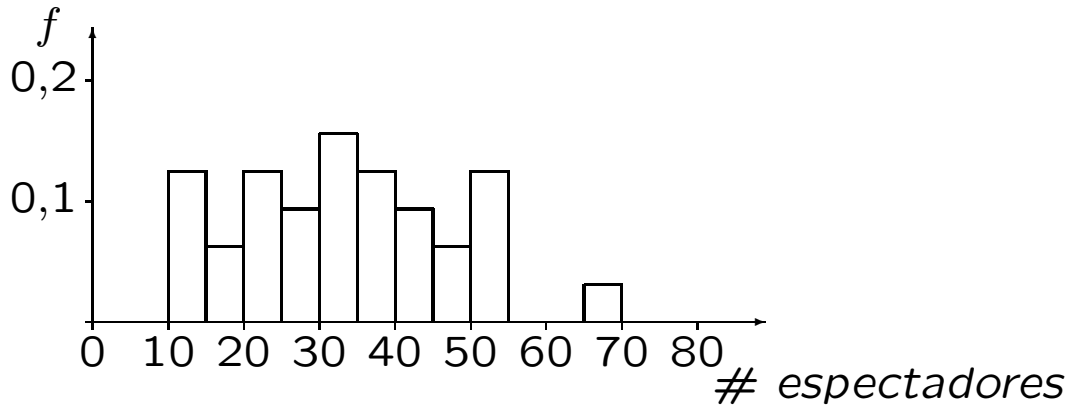


Con diferencia a los datos discretas, las barras están conectadas.

Además, observamos que igual que para los datos discretas, es posible construir el histograma con frecuencias absolutas o relativas o con frecuencias (absolutas o relativas) acumuladas.

¿Cómo elegir el número de barras?

Ejemplo 23 *En el Ejemplo 22 vemos que pasa si usamos 14 intervalos* $[10, 15)$, $[15, 20)$, ...



Con demasiadas barras (o muy pocas barras), se pierde un poco la idea de la forma de la distribución. ¡Con sólo una barra es aún peor! Una **regla empírica** razonable es elegir aproximadamente $\sqrt{n}$ barras donde n es el tamaño de la muestra.

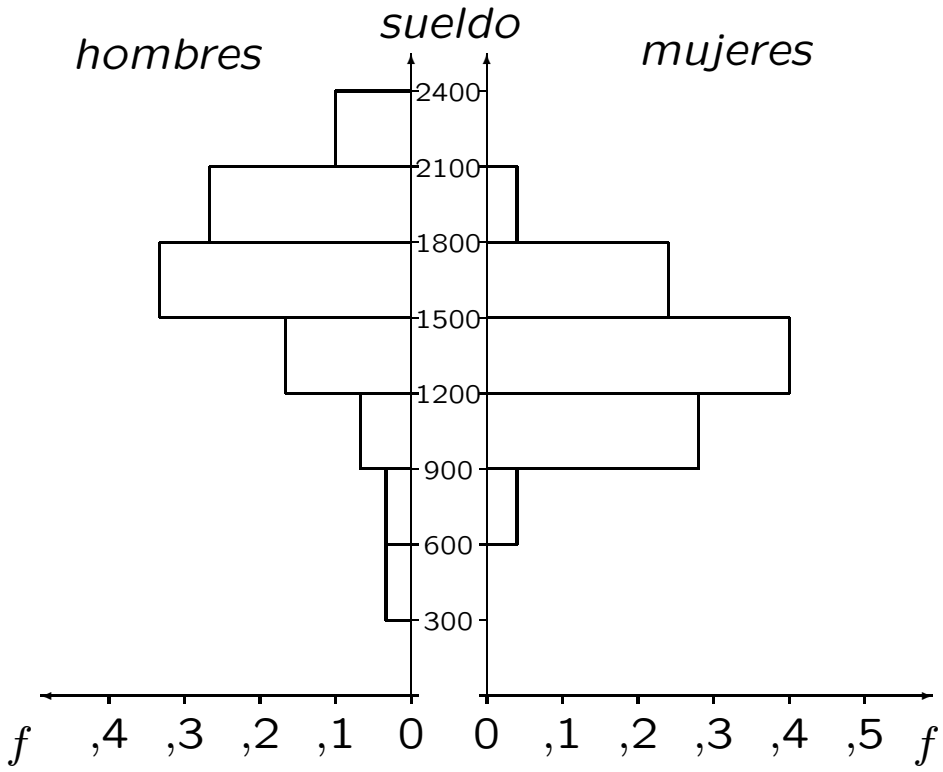
Histogramas para comparar dos grupos

Ejemplo 24 *La tabla resume las ganancias por hora de una muestra de 30 hombres y 25 mujeres, con estudios secondarios (o más) y trabajando $>$ 15 horas semanales en españa.*

	H		*M*	
Intervalo	n_i	f_i	n_i	f_i
$[300, 600)$	1	,033	0	0
$[600, 900)$	1	,033	1	,04
$[900, 1200)$	2	,067	7	,28
$[1200, 1500)$	5	,167	10	,4
$[1500, 1800)$	10	,333	6	,24
$[1800, 2100)$	8	,267	1	,04
$[2100, 2400)$	3	,100	0	0
> 2400	0	0	0	0
	30	1	25	1

Usamos dos histogramas con la misma escala para representar los datos.

Histograma de los sueldos horarios de hombres y mujeres



El sueldo medio de los hombres parece un poco más alto y la distribución de sueldo de hombres es más dispersa y asimétrica.

Dolado, J. y V. LLorens (2004). Gender Wage Gaps by Education in Spain: Glass Floors vs. Glass Ceilings, *CEPR DP.*, **4203**.

`http://www.eco.uc3m.es/temp/dollorens2.pdf`

Histogramos con intervalos de distíntos tamaños

En este caso, se construyen las barras para que el área de cada barra es proporcional al número de datos.

Ejemplo 25 *Los siguientes datos son los resultados de una encuesta de usuarios sobre el número de gramas de marijuana que fuman cada semana.*

g / semana	*Frecuencia*
$[0, 3)$	94
$[3, 11)$	269
$[11, 18)$	70
$[18, 25)$	48
$[25, 32)$	31
$[32, 39)$	10
$[39, 46)$	5
$[46, 74)$	2
> 74	0

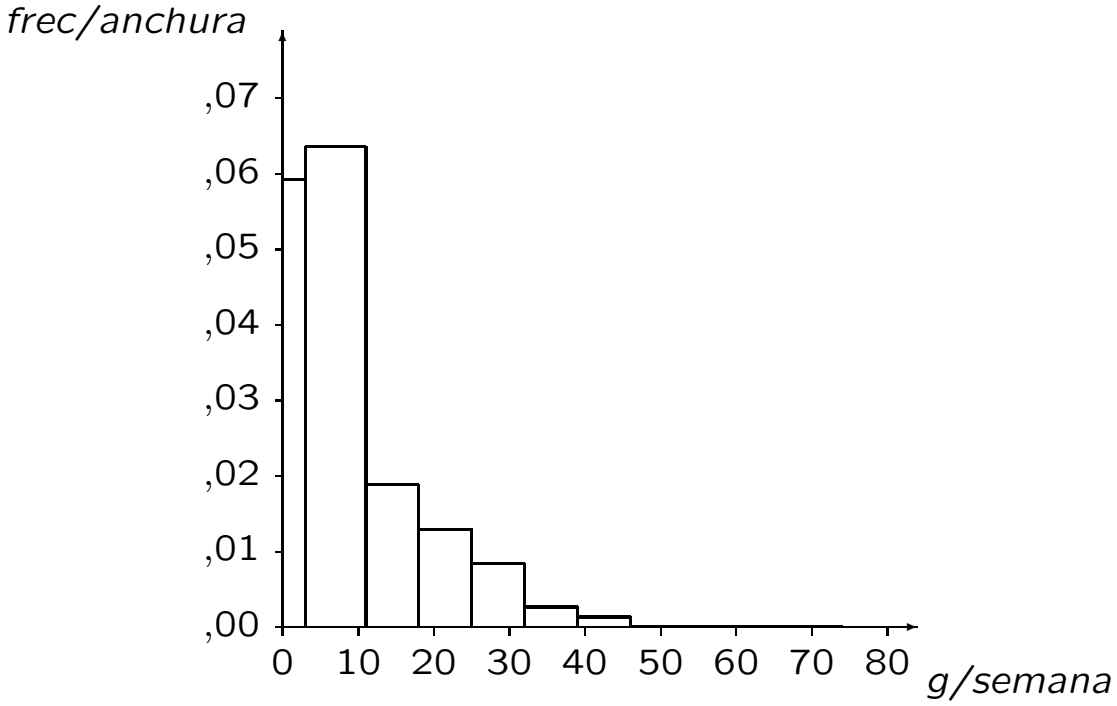
Aumentamos la tabla con las frecuencias relativas y las alturas de las barras.

g / semana	*anchura*	n_i	f_i	*altura*
$[0, 3)$	3	94	,178	,0592
$[3, 11)$	8	269	,509	,0636
$[11, 18)$	7	70	,132	,0189
$[18, 25)$	7	48	,091	,0130
$[25, 32)$	7	31	,059	,0084
$[32, 39)$	7	10	,019	,0027
$[39, 46)$	7	5	,009	,0014
$[46, 74)$	28	2	,004	,0001
> 74	0	0	0	0
Total		529	1	

Usamos la fórmula

altura = frecuencia/anchura del intervalo

Histograma del consumo de marijuana semanal



Se ve claramente que la distribución es muy asimétrica a la derecha.

Landrigan et al (1983). Paraquat and marijuana: epidemiologic risk assessment. *Amer. J. Public Health*, **73**, 784-788

Otros gráficos

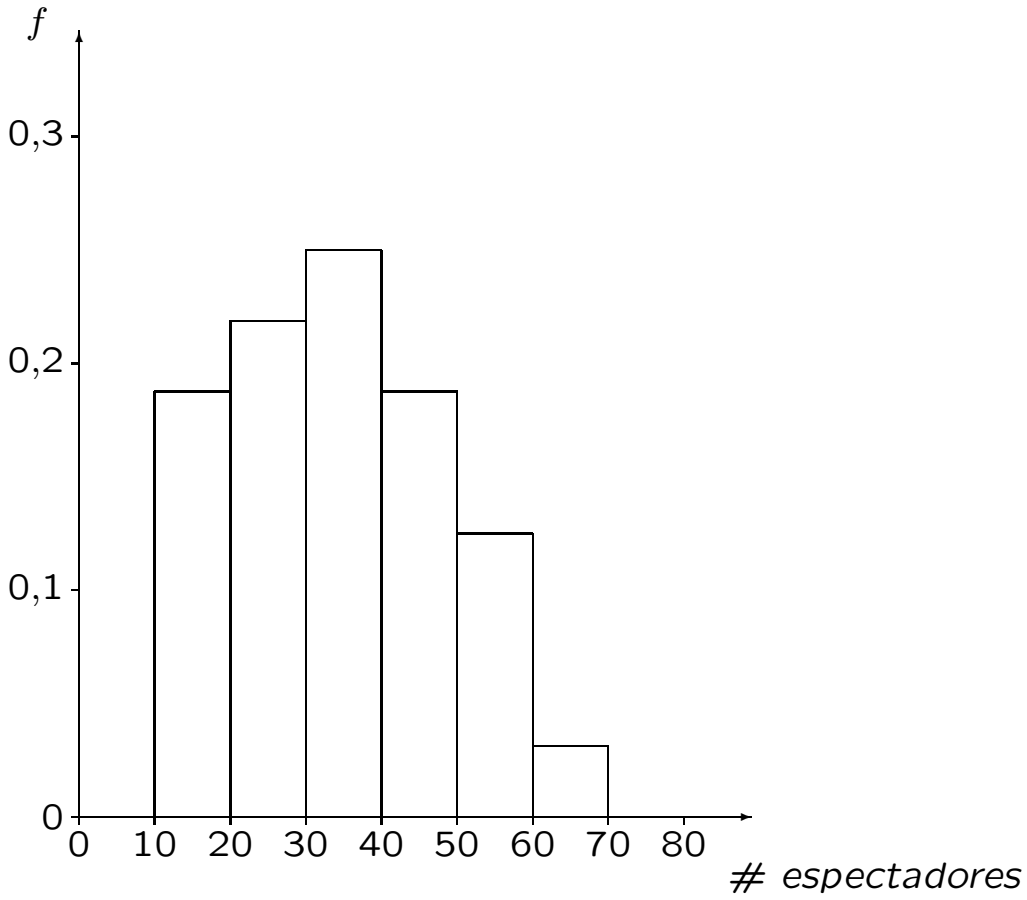
Existen varias alternativas al histograma.

1) El polígono de frecuencias

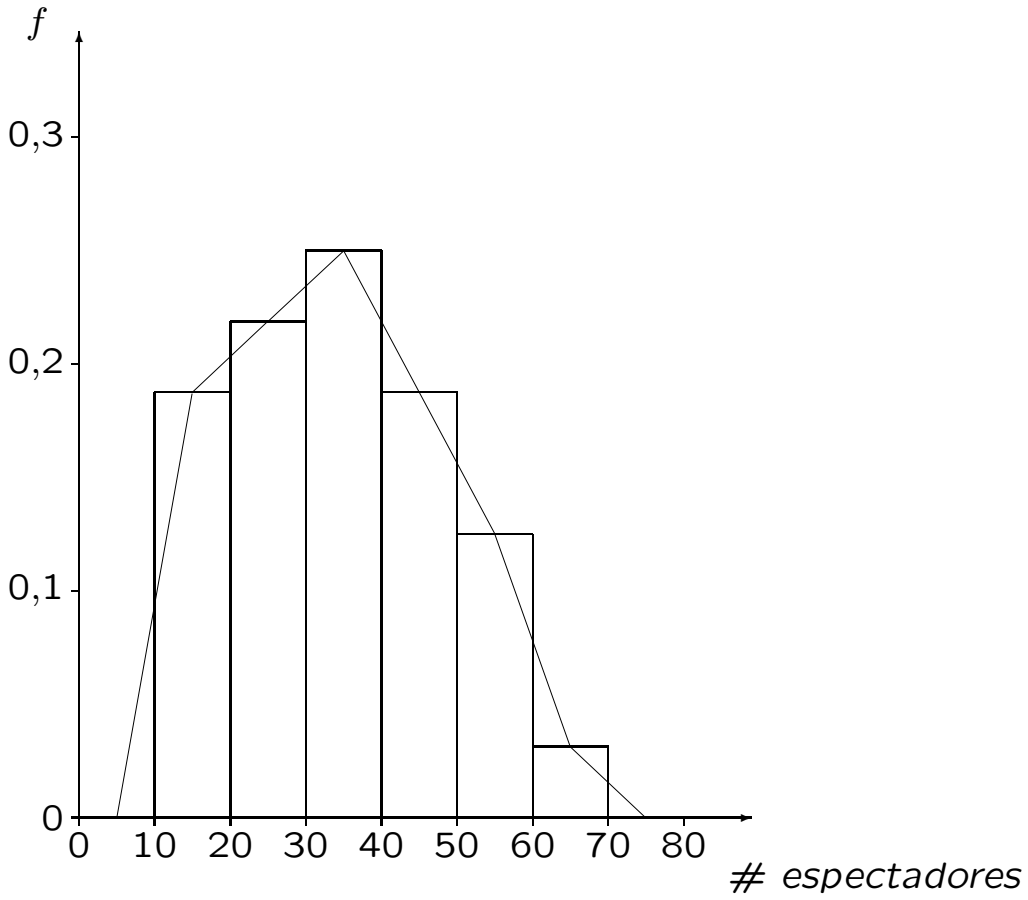
Es basícamente una versión suavizado del histograma de frecuencias relativas.

Ejemplo 26 *Retomamos el Ejemplo 22, y construimos un polígono de frecuencias relativas.*

Empezamos con el histograma, ...



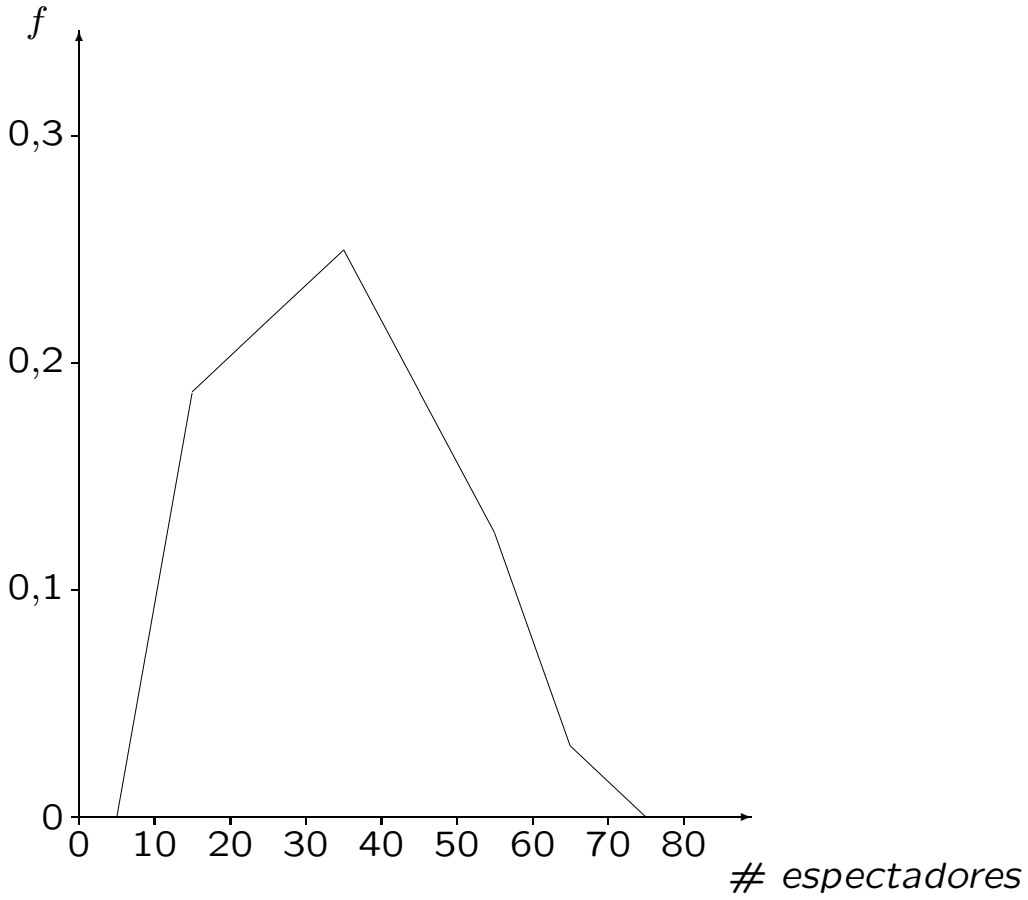
y unimos los centros de cada barra.



Se une el polígono de frecuencias al eje x en el centro de un intervalo vacio a cada lado del histograma.

Por último, se quita el histogram

Polígono de frecuencias



El polígono de frecuencias acumuladas

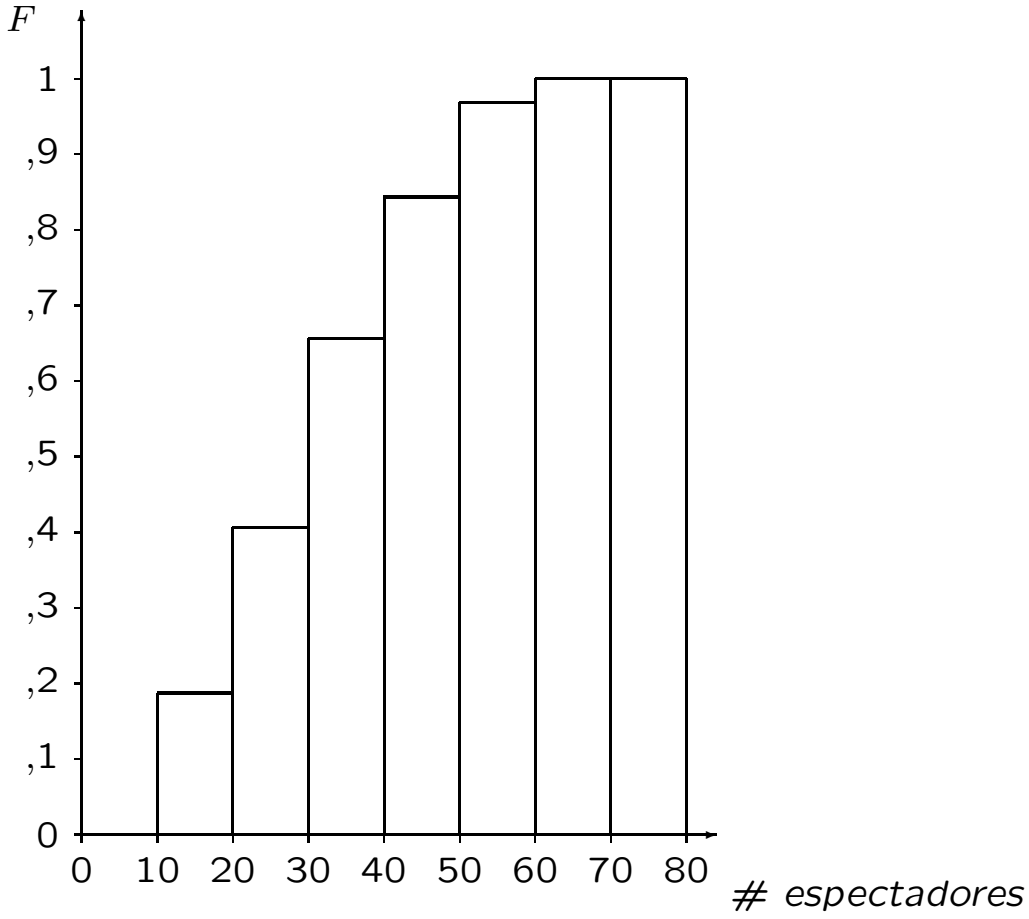
Se construye un polígono a través del histograma de frecuencias acumuladas.

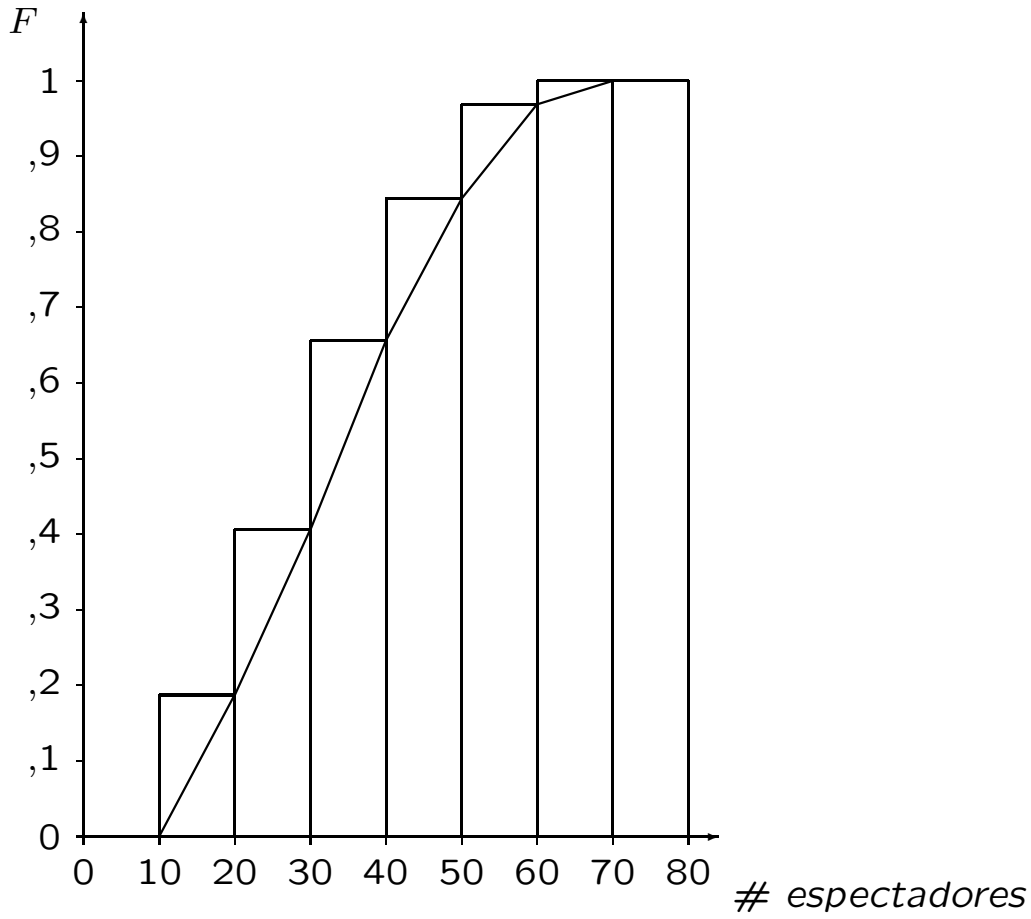
Ejemplo 27 *En el Ejemplo 22, construimos una tabla de frecuencias relativas acumuladas.*

Clase	n_i	f_i	F_i
$[10, 20)$	6	,18750	,18750
$[20, 30)$	7	,21875	,40625
$[30, 40)$	8	,25000	,65625
$[40, 50)$	6	,18750	,84375
$[50, 60)$	4	,12500	,96875
$[60, 70)$	1	,03125	1,00000
> 70	0	,00000	1,00000
Total	32	1,00000	

El gráfico es un histograma de frecuencias relativas acumuladas.

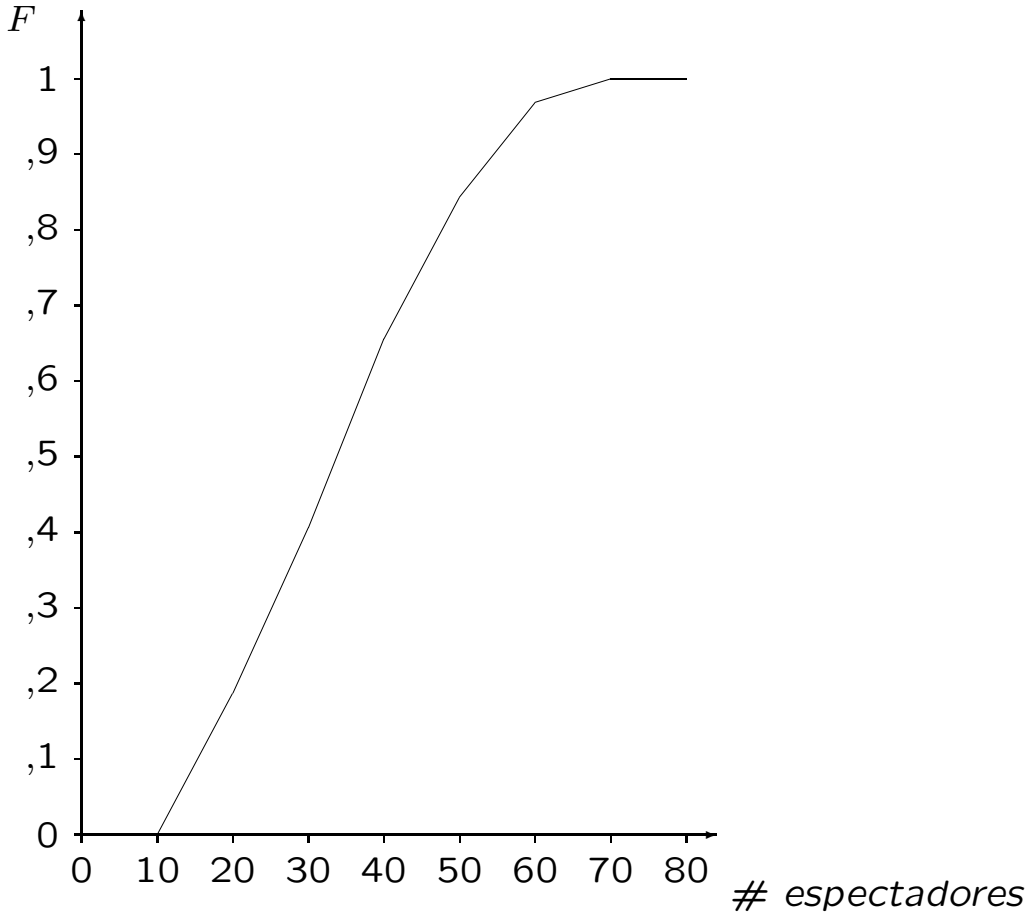
Histograma de frecuencias acumuladas





Para construir el polígono, se unen las rectas a la derecha de cada intervalo y finalmente, se quita el histograma.

Polígono de frecuencias acumuladas



2) El diagrama de tallo y hojas

Es parecido a un histograma pero preserva los valores numéricos de los datos originales.

Ejemplo 28 *Los siguientes datos son los resultados de 26 alumnos en una prueba de estadística.*

64	82	85	99	96	81	97
80	81	80	84	87	98	75
86	88	82	78	81	86	80
50	84	88	83	82		

El mínimo es 50 *y el máximo es* 99. *Dividimos los datos en unidades y décimos.*

5	0																	
6	4																	
7	5	8																
8	2	5	1	0	1	0	4	7	6	8	2	1	6	0	4	8	3	2
9	9	6	7	8														

Ahora, ordenamos las hojas y añadimos una clave explicatoria.

Diagrama de tallo y hojas de los resultados de una prueba de estadística

5	0
6	4
7	5 8
8	0 0 0 1 1 1 2 2 2 3 4 4 5 6 6 7 8 8
9	6 7 8 9

Tallo		Décimos
Hoja		Unidades
1|1	=	11

En algunos ejemplos, es necesario dividir los intervalos.

Ejemplo 29 *Los datos son los emisiones de dióxido de 47 estados americanos (lb/millón Btu).*

0,3	2,3	4,2	3,8	1,5	0,6	0,4	0,5	1,5	1,3
4,5	3,6	1,2	1,2	3,4	0,2	0,7	0,2	0,7	4,1
1,0	2,7	2,2	2,5	2,7	1,7	1,5	3,7	2,9	1,5
2,1	1,5	1,4	1,9	1,0	2,9	1,7	1,8	1,7	0,6
0,9	0,6	1,8	1,4	2,0	2,1	3,5			

Usamos las unidades para el tallo y los decimales para las hojas.

Datos basados en Friedman et al (1983). *The American Statistician*, **37**, 385-394.

En primer lugar, supongamos que usamos barras de anchura 1*.*

0	3	6	4	5	2	7	2	7	6	9	6							
1	5	5	3	2	2	0	7	5	5	5	4	9	0	7	8	7	8	4
2	3	7	2	5	7	9	1	9	0	1								
3	8	6	4	7	5													
4	2	5	1															

El diagrama tiene pocas hojas. Mejor es dividir los intervalos. Usamos intervalos de tamaño 0,5*. (La otra posibilidad sería intervalos de tamaño* 0,2*, pero serían demasiados.)*

0	3	4	2	2						
0	6	5	7	7	6	9	6			
1	3	2	2	0	4	0	4			
1	5	5	7	5	5	5	9	7	8	7	8
2	3	2	1	0	1					
2	7	5	7	9	9					
3	4									
3	8	6	7	5						
4	2	1								
4	5									

Ya se ve bien la forma de la distribución. Finalmente, ordenamos las hojas y añadimos una clave explicatoria.

Diagrama de tallo y hojas de emisiones de dióxido

0	2	2	3	4							
0	5	6	6	6	7	7	9				
1	0	0	2	2	3	4	4				
1	5	5	5	5	5	7	7	7	8	8	9
2	0	1	1	2	3						
2	5	7	7	9	9						
3	4										
3	5	6	7	8							
4	1	2									
4	5										

Tallo		*Unidades*
Hoja		*Decimales*
1|1	=	1,1

En otros ejemplos, es necesario usar 2 dígitos enla hoja para cada número. Sino, se pierde información.

Ejemplo 30 *Volvemos al Ejemplo 22. En este caso, usamos décimos para la talla y representamos los números con 2 dígitos en la hoja. Recordamos ordenar las hojas.*

1	03 09 22 43 61 78
2	11 30 37 52 90 91 93
3	00 13 17 23 42 52 52 54 61
4	21 32 60 70 72
5	10 14 16 17
6	70

Tallo		*Décimos*
Hoja		*Decimales*
1|23	=	12,3

Ejemplo 31 *Vemos el diagrama de tallo y hojas hecho en Statgraphics.*

```
Stem-and-Leaf Display for Espec: unit = 1,0
1|2 represents 12,0

      4     1|0024
      6     1|67
      9     2|133
     13     2|5999
    (5)     3|01124
     14     3|5556
     10     4|23
      8     4|677
      5     5|1111
      1     5|
      1     6|
      1     6|7
```

Se pierden los decimales.