1. LA ESTADÍSTICA DESCRIPTIVA

Objetivo

Aprender cómo resumir las características más importantes de una muestra de datos.

Bibliografia recomendada

Peña y Romo (1997), Capítulos 1 – 5.

Newbold (1997) Capítulos 1 – 2.

Índice

Conceptos fundamentales

Definición 4 *La* **población** *es el conjunto de individuos o elementos, que se quiere estudiar.*

Ejemplo 6

i) *La población de gente en España.*

ii) *Los donantes de sangre en España.*

iii) *La población de asientos en el estadio Santiago Bernabeu en el siguiente partido frente a Barça.*

iv) *Los diabéticos en Madrid.*

Observamos que una población puede ser tanto finita como infinita.

Definición 5 *El fenómeno o característica de la población que se quiere estudiar se llama* **una variable**.

Ejemplo 7 *Retomando el Ejemplo 6:*

i) *La edad en años. Posibles valores* $\{0, 1, 2, \ldots\}$

ii) *El tipo de sangre.* $\{A, B, AB, O\}$

iii) *La cantidad de dinero pagado para un asiento.* $[0, \infty)$

iv) *Nivel de azúcar en la sangre.* $\{$*alto, mediano, bajo*$\}$

Tipos de datos

Es posible clasificar distintos tipos de variables. En primer lugar, se distinguen entre las variables de naturaleza categorica y las variables de naturaleza numérica.

Definición 6 *Una variable* **cualitativa** *o atributo es una variable que no aparece en forma numérica, sino como categorías o atributos.*

Ejemplo 8 *En el Ejemplo 7, el tipo de sangre o el nivel de azúcar en el cuerpo son variables cualitativas.*

Definición 7 *Una variable* **cuantitativa** *es una variable que puede expresarse numéricamente.*

Ejemplo 9 *En el Ejemplo 7, la edad y el precio del asiento son variables cuantitativas.*

Las variables cualitativas se dividen en variables **nominales** y variables **ordinales**. Son nominales si las distíntas clases no tienen una orden natural y son ordinales si las categorías están ordenadas.

Ejemplo 10 *Volviendo al Ejemplo 7, el tipo de sangre es una variable nominal y el nivel de azúcar es ordinal.*

Igualmente, se dividen las variables cuantitativas en dos clases: variables **discretas** y variables **continuas**. Una variable discreta es una variable que puede tomar una clase fija de distíntos valores. Una variable continua puede tomar cualquier valor en un rango continuo.

Ejemplo 11 *En el Ejemplo 7, la edad es una variable discreta y el precio del asiento es continua.*

Cómo resumir una muestra de datos cualitativos

Dada una muestra de datos, se quiere extraer la información pertinente. Mostrando a alguien la muestra entera, no van a ser capaces de ver los rasgos importantes.

Ejemplo 12 *Se quería estudiar los niveles de educación de la gente en Getafe y se preguntó a 50 personas sus niveles de estudios (***Si***n Estudios,* **P***rimario,* **S***econdario,* **B***achillerato,* **U***niversitarios) con los siguientes resultados.*

U	*B*	*U*	*S*	*S*	*P*	*P*	*Si*	*B*	*B*
S	*U*	*B*	*B*	*B*	*S*	*P*	*S*	*B*	*B*
Si	*P*	*P*	*P*	*S*	*U*	*B*	*B*	*B*	*S*
U	*U*	*S*	*B*	*S*	*S*	*B*	*B*	*P*	*S*
S	*B*	*B*	*S*	*B*	*P*	*S*	*B*	*S*	*B*

¿Qué tipo de variable es el nivel de estudios?

La tabla de frecuencias

Es muy difícil distinguir cúal es el nivel de estudios más típico. Para resumir los datos, en primer lugar, es conveniente hacer una tabla de las frecuencias en cada categoría.

Ejemplo 13 *Volvemos al Ejemplo 12.*

Categoría	*Frecuencia absoluta*
Si	*2*
P	*8*
S	*15*
B	*19*
U	*6*
Total	*50*

Observamos que se ordena la tabla desde el nivel de estudios más bajo hasta el nivel más alto.

Se ve que la clase más frecuente es de estudios secondarios. Supongamos que se interesaba por la proporción de gente sin estudios de secondario.

Se aumenta la tabla con **frecuencias relativas**

Categoría	*Frecuencia absoluta*	*Frecuencia relativa*
Si	*2*	0,04
P	*8*	0,16
S	*15*	0,30
B	*19*	0,38
U	*6*	0,12
Total	*50*	1

Entonces, un $(0{,}04 + 0{,}16) \times 100\,\% = 20\,\%$ *de la gente en la muestra no tenía estudios secondarios.*

Cómo construir una tabla de frecuencias

Para una variable con k categorías $C_1, \ldots, C_k$, sea n_k el número de observaciones en cada clase. Entonces se construye la tabla de frecuencias.

Categoría	Frecuencia absoluta	Frecuencia relativa
C_1	n_1	$f_1 = \frac{n_1}{n}$
C_2	n_2	$f_2 = \frac{n_2}{n}$
$\vdots$	$\vdots$	$\vdots$
C_k	n_k	$f_k = \frac{n_k}{n}$
Total	$n = n_1 + n_2 + \ldots + n_k$	1

La tabla que presenta las clases o categorías de las variables y sus respectivas frecuencias se llama **la distribución de frecuencias**.

Ejemplo 14 *El Departamento de Estadística tiene interés en las licenciaturas (***Ec***onomía, Economía de la* **Emp***resa o Estudios* **C***onjuntos) que cursan los alumnos de Introducción a la Estadística. Toman una muestra aleatoria de 40 estudiantes con los siguientes resultados.*

Ec	*Emp*	*Ec*	*Ec*	*C*	*Emp*	*Emp*	*Ec*
C	*Ec*	*Ec*	*Emp*	*Emp*	*Emp*	*Ec*	*Ec*
Ec	*Ec*	*C*	*C*	*Emp*	*Emp*	*Emp*	*Ec*
C	*C*	*Ec*	*Ec*	*Emp*	*Emp*	*Ec*	*Emp*
Ec	*Ec*	*Emp*	*Emp*	*C*	*Ec*	*Ec*	*Emp*

Esta variable es nominal. Entonces para construir la tabla de frecuencias, el orden no importa. Es normal ordenar las categorias alfabéticamente.

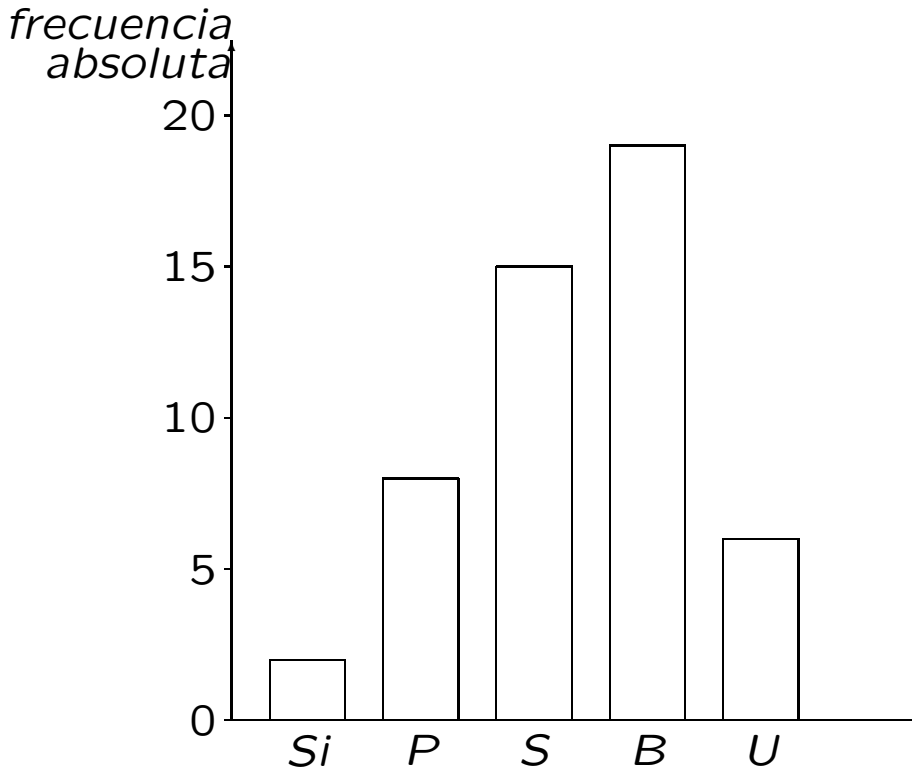
Categoría	*Frecuencia absoluta*	*Frecuencia relativa*
C	*7*	0,175
Ec	*18*	0,450
Emp	*15*	0,375
Total	*40*	1

El diagrama de barras

La gente prefiere ver imagenes que tablas de números y entonces es útil usar gráficos para mostrar los datos. El gráfico más importante para variables cualitativas es el **diagrama de barras**.

Ejemplo 15 *Construimos un diagrama de barras de los datos del Ejemplo 12 sobre estudios.*

Diagrama de barras de los niveles de estudios de los Getafenses



Clave

Si = *Sin Estudios*
P = *Primario*
S = *Secondario*
B = *Bachillerato*
U = *Universitarios*

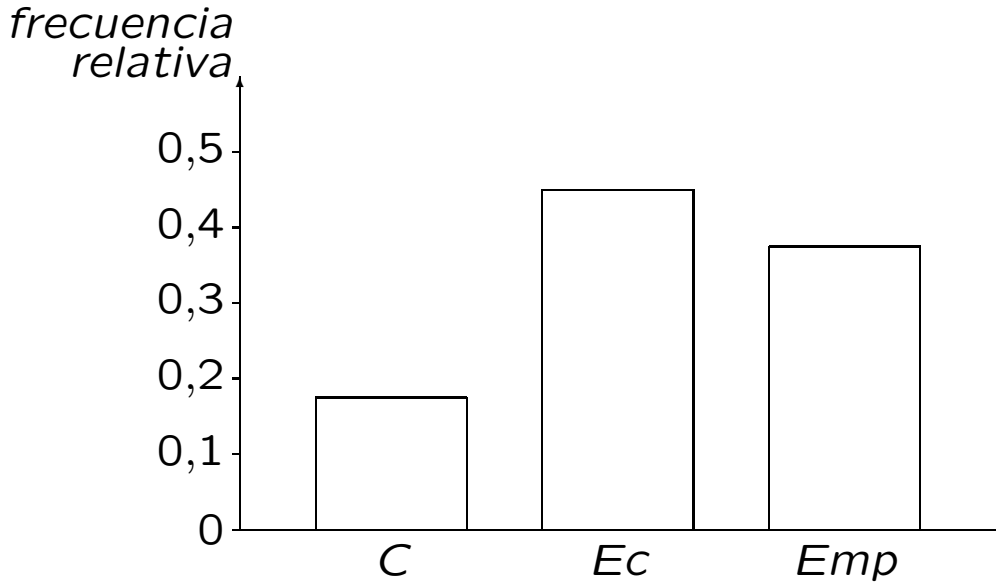
Es importante observar que como la variable en este ejemplo es ordinal, es natural ordenar las barras en el orden de las categorías de la variable desde la más baja (Sin estudios) hasta la más alta (Universitaria).

Si la variable es nominal, el orden de las barras no importa tanto. Lo más natural es ordenar las barras alfabéticamente.

También es posible construir un diagrama de barras usando frecuencias relativas en lugar de frecuencias absolutas.

Ejemplo 16 *Construimos un diagrama de barras para los datos del Ejemplo 14.*

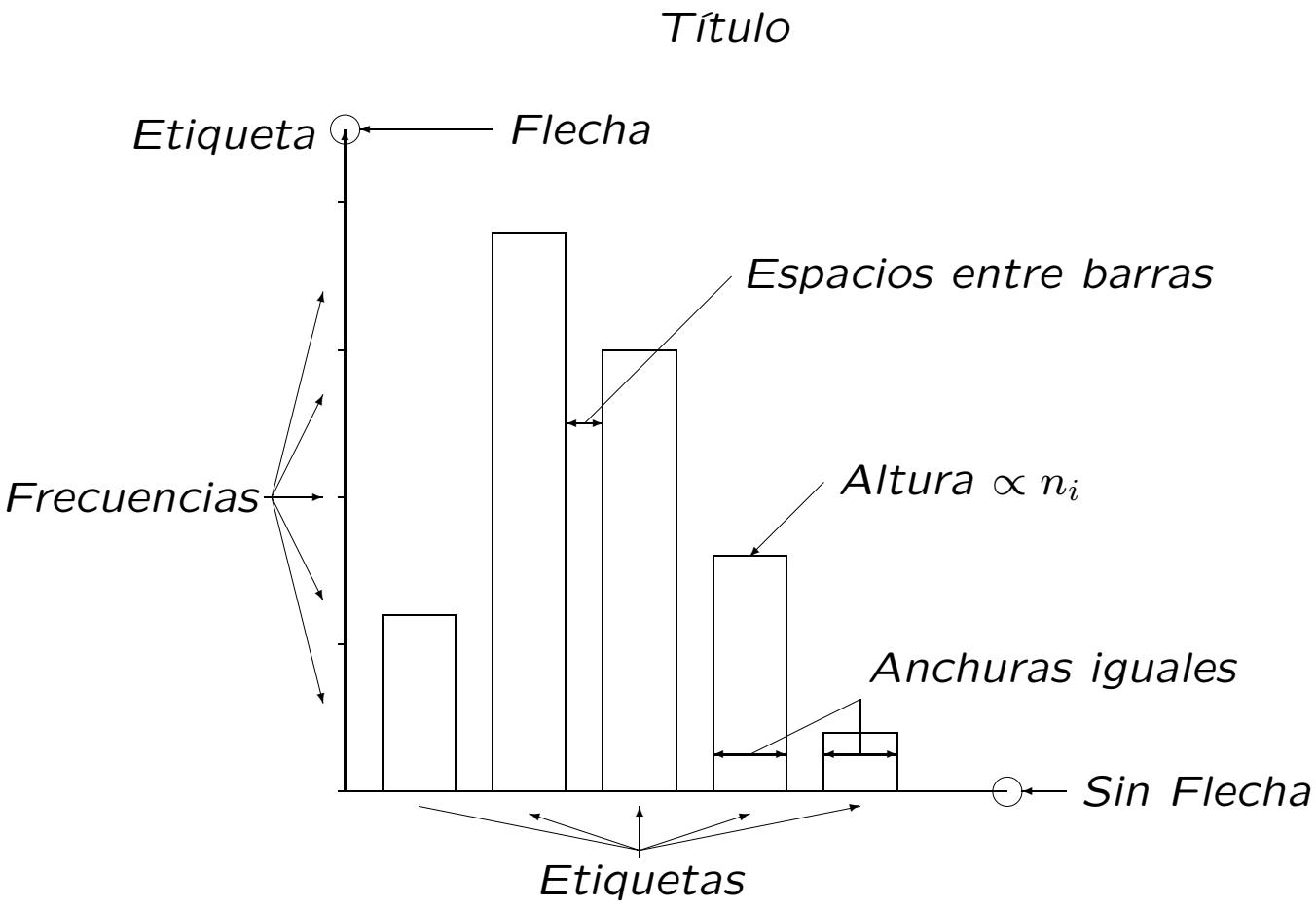
Programas de licenciatura de estudiantes de Estadística



Clave

$C =$ *Estudios Conjuntos*
$Ec =$ *Economía*
$Emp =$ *Economía de la Empresa*

Los rasgos importantes de un diagrama de barras



Clave

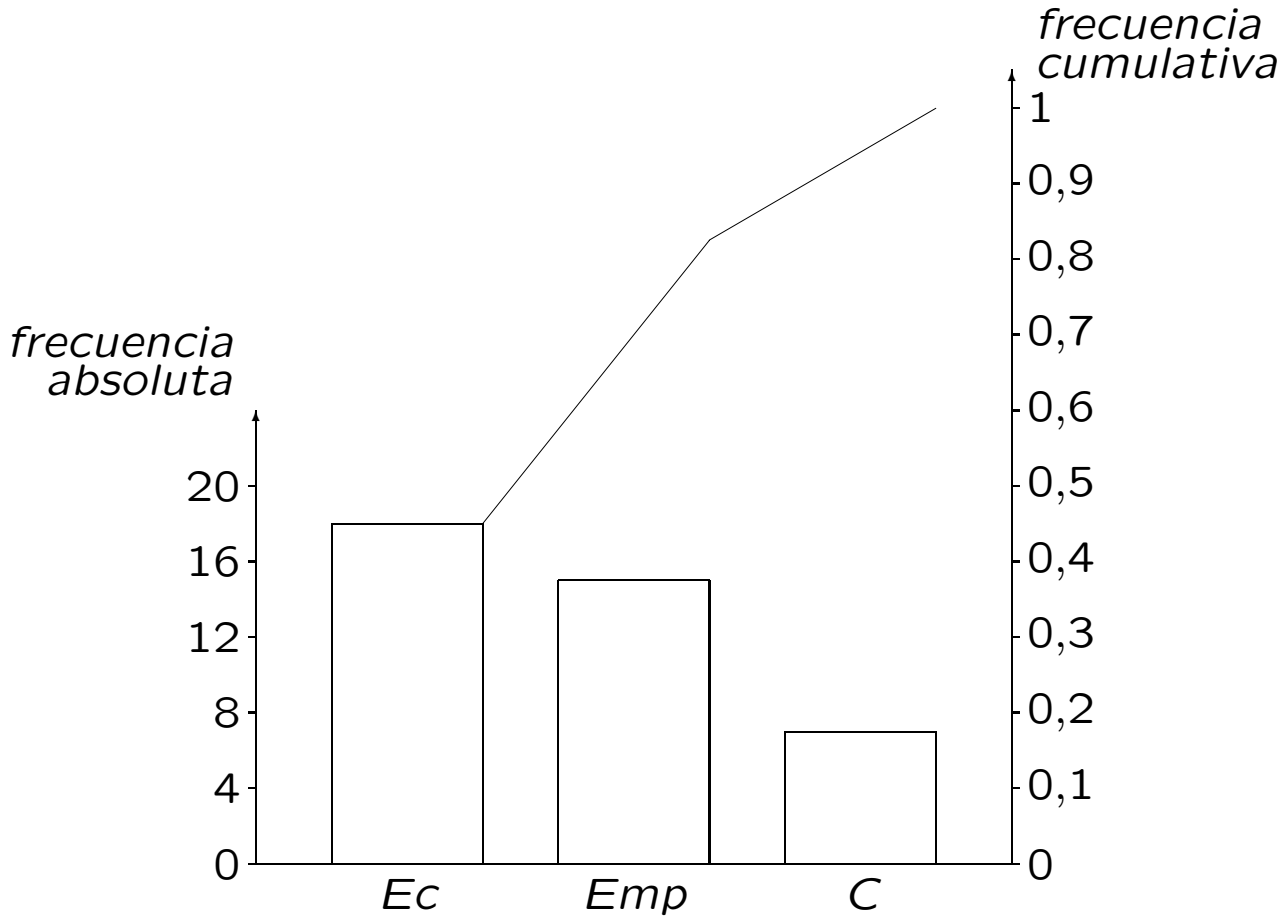
Otros gráficos para datos cualitativos

Si se ordenan las categorías de más a menos frecuentes y se dibuja un diagrama de barras de frecuencias absolutas, añadiendo una linea para mostrar las frecuencias relativas acumuladas, se tiene un **diagrama de Pareto**.

Ejemplo 17 *Retomamos el Ejemplo anterior. Ordenamos las categorías en términos de frecuencia y calculamos las frecuencias acumuladas.*

Categoría	*Frecuencia absoluta*	*Frecuencia relativa*	*Frecuencia acumulada*
Ec	*18*	0,450	0,450
Emp	*15*	0,375	0,825
C	*7*	0,175	1
Total	*40*	1	—

Diagrama de Pareto de programas de licenciatura



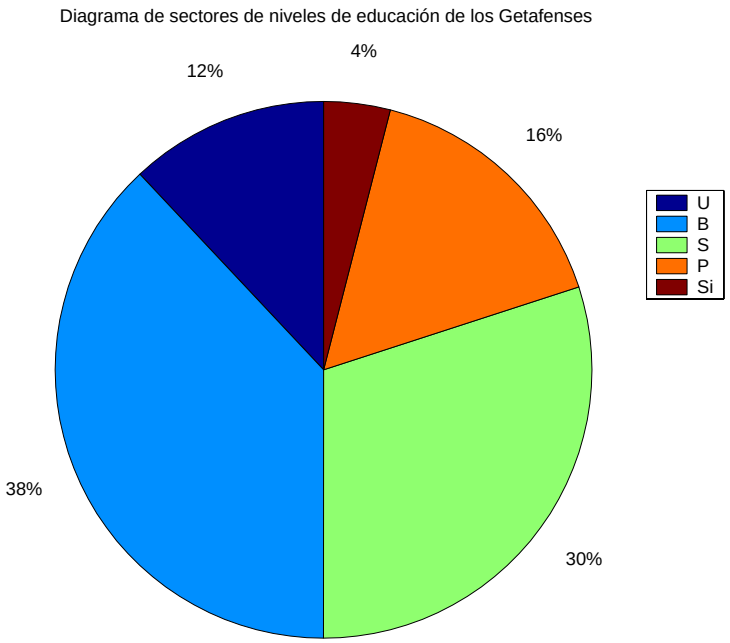
Clave

Ec = *Economía*
Emp = *Economía de la Empresa*
C = *Estudios Conjuntos*

El diagrama de sectores o de pastel

Se divide un círculo en sectores donde el área de un sector es proporcional al número de datos en una categoría.

Ejemplo 18 *Se ilustra un diagrama de sectores de los datos del nivel de educación de los Getafenses.*



Pictogramas etc.

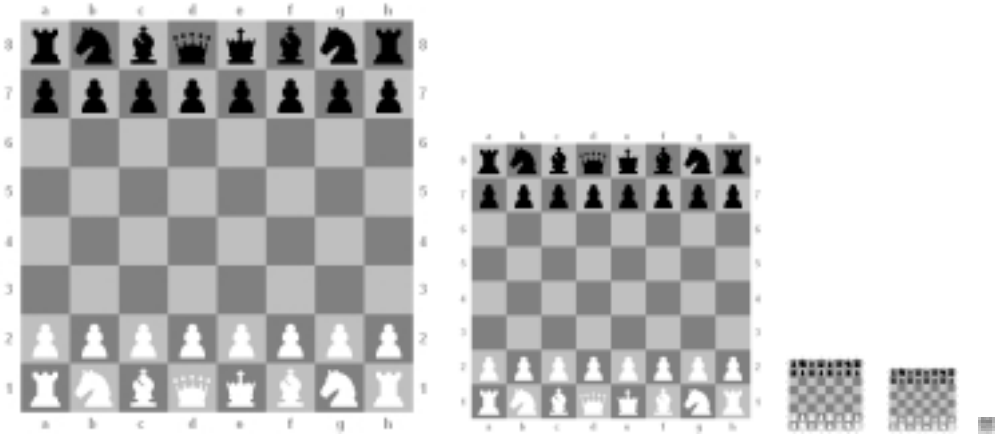
Un pictograma es una representación gráfica usando dibujos relevantes para ilustrar los datos, en lugar de simples barras. Son de muchos estilos y formas.

Ejemplo 19 *La tabla muestra las frecuencias de las distíntas primeras jugadas encontradas en el 20 de febrero del 2006 en el buscador de aperturas de* `http://www.chessgames.com/`.

Apertura	*Frecuencia*
$e4$	178130
$d4$	125919
$Cf3$	32206
$c4$	28796
$Otras$	6480
$Total$	371631

Usamos tableros con tamaños proporcionales a frecuencias *para representar los datos.*

Pictograma de las aperturas utilizadas en el ajedrez



e4 *d4* *Cf3* *c4* *Otr.*

El diagrama de barras con datos discretas

Es posible construir un diagrama de barras para una muestra de datos discretas.

Ejemplo 20 *Un estadístico decidió grabar el número de cartas que recibió durante 30 días laborales con los siguientes resultados.*

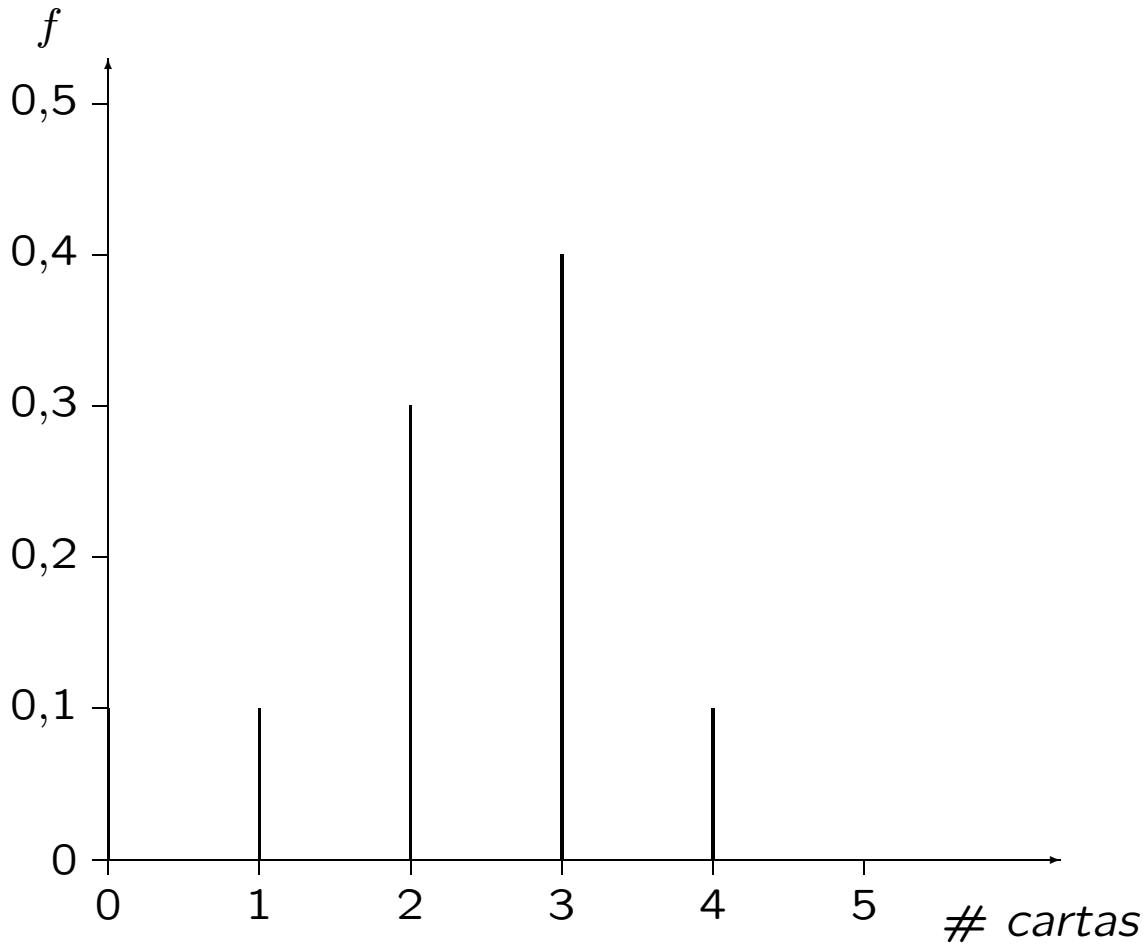
2	3	2	3	3	1	0	4	2	3
3	2	0	3	3	3	2	3	4	1
2	2	1	2	3	0	3	2	4	3

En primer lugar, construimos una tabla de frecuencias de estos datos.

Número	*Frecuencia absoluta*	*Frecuencia relativa*
0	3	0,1
1	3	0,1
2	9	0,3
3	12	0,4
4	3	0,1
> 4	0	0
Total	30	1

Hemos incluido una fila vacia (> 4).

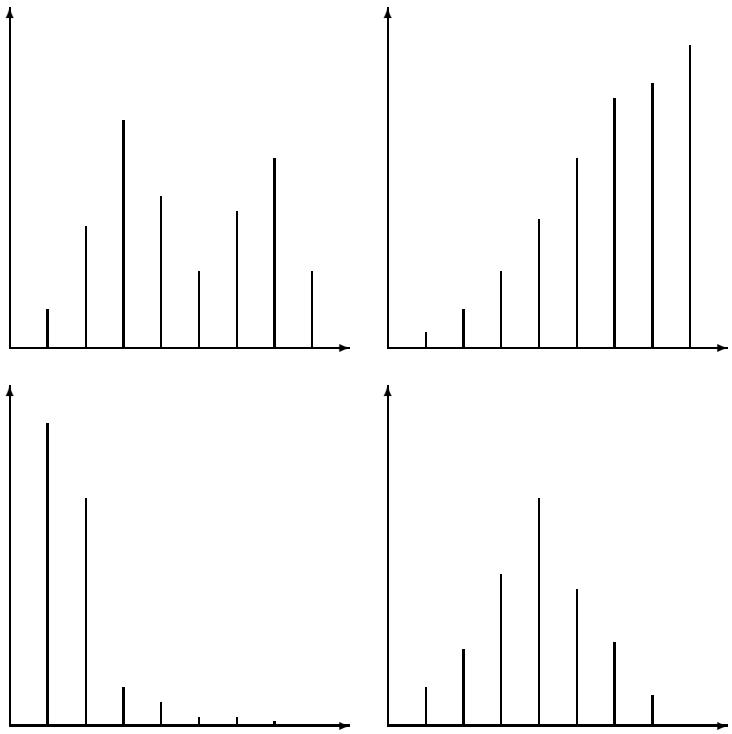
Diagrama de barras de cartas recibidas por día



Es habitual poner rectas (o barras muy finas) en lugar de las barras anchas usadas en los ejemplos anteriores. Además, incluimos una clase vacia al final.

Rasgos de los datos

En el ejemplo anterior, podemos ver que la distribución de los datos es **unimodal**, (con una moda de 3 cartas por día) y un poco **asimétrica a la izquierda**.



Frecuencias acumuladas

Ejemplo 21 *Volviendo al Ejemplo 20, puede que el estadístico tenga interés en la proporción de días en los cuales ha recibido menos de dos cartas.*

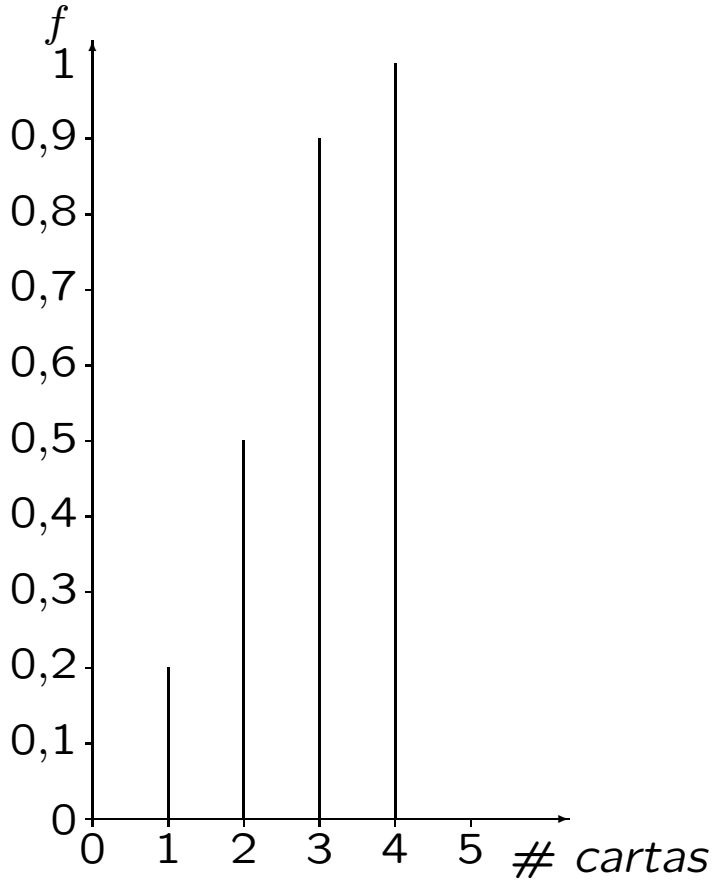
Podemos calcular el resultado mediante una tabla de frecuencias acumuladas.

Número	*Frec. abs.*	*Frec. rel.*	*Frec. abs. acumulada*	*Frec. rel. acumulada*
0	3	0,1	3	0,1
1	3	0,1	6	0,2
2	9	0,3	15	0,5
3	12	0,4	27	0,9
4	3	0,1	30	1
> 4	0	0	30	1

La proporción de días es un 20 %.

Además, podemos hacer un diagrama para mostrar la distribución de frecuencias acumuladas.

Frecuencias acumuladas de cartas recibidas por día



Método general para construir una tabla de frecuencias acumuladas

Supongamos que tenemos una muestra de datos $x_1 < \ldots < x_k$ con frecuencias absolutas $n_1, \ldots, n_k$.

i	x_i	n_i	f_i	N_i	F_i
1	x_1	n_1	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
2	x_2	n_2	$f_2 = \frac{n_2}{n}$	$N_2 = n_1 + n_2$	$F_1 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_k	n_k	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
$> k$	$> x_k$	0	0	n	1
		n	1		