



**FACULTAD DE INGENIERIA U.N.A.M.  
DIVISION DE EDUCACION CONTINUA**

**A LOS ASISTENTES A LOS CURSOS**

**Las autoridades de la Facultad de Ingeniería, por conducto del jefe de la División de Educación Continua, otorgan una constancia de asistencia a quienes cumplan con los requisitos establecidos para cada curso.**

**El control de asistencia se llevará a cabo a través de la persona que le entregó las notas. Las inasistencias serán computadas por las autoridades de la División, con el fin de entregarle constancia solamente a los alumnos que tengan un mínimo de 80% de asistencias.**

**Pedimos a los asistentes recoger su constancia el día de la clausura. Estas se retendrán por el periodo de un año, pasado este tiempo la DECFI no se hará responsable de este documento.**

**Se recomienda a los asistentes participar activamente con sus ideas y experiencias, pues los cursos que ofrece la División están planeados para que los profesores expongan una tesis, pero sobre todo, para que coordinen las opiniones de todos los interesados, constituyendo verdaderos seminarios.**

**Es muy importante que todos los asistentes llenen y entreguen su hoja de inscripción al inicio del curso, información que servirá para integrar un directorio de asistentes, que se entregará oportunamente.**

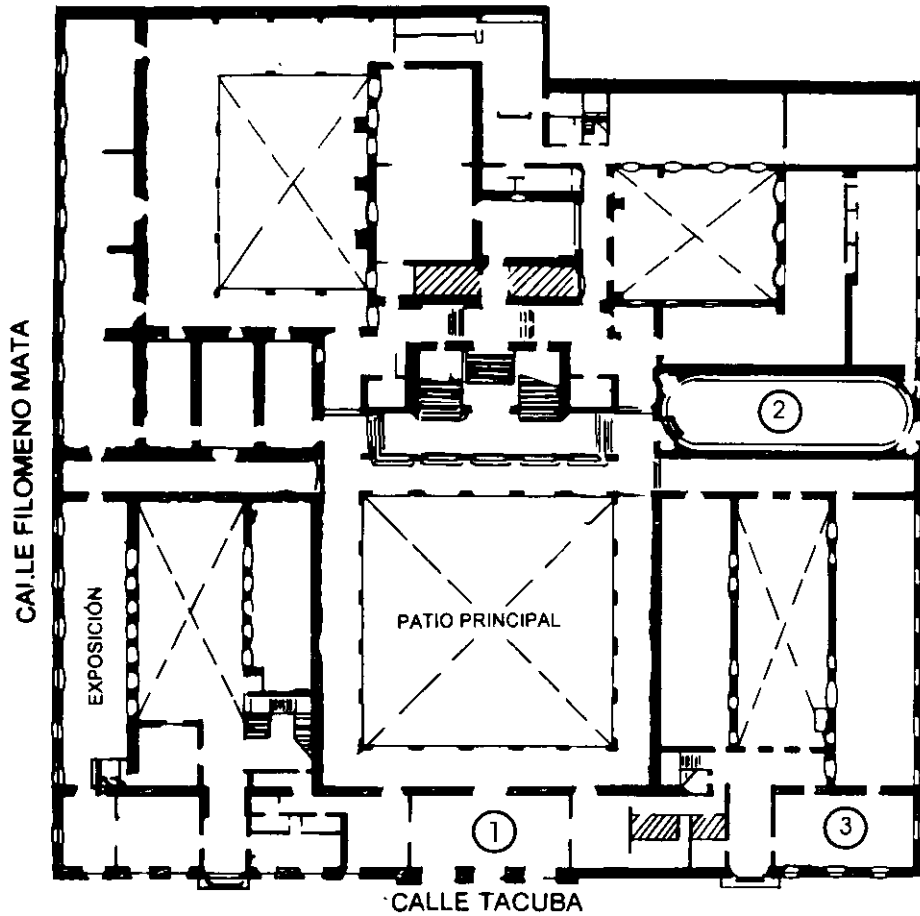
**Con el objeto de mejorar los servicios que la División de Educación Continua ofrece, al final del curso deberán entregar la evaluación a través de un cuestionario diseñado para emitir juicios anónimos.**

**Se recomienda llenar dicha evaluación conforme los profesores impartan sus clases, a efecto de no llenar en la última sesión las evaluaciones y con esto sean más fehacientes sus apreciaciones.**

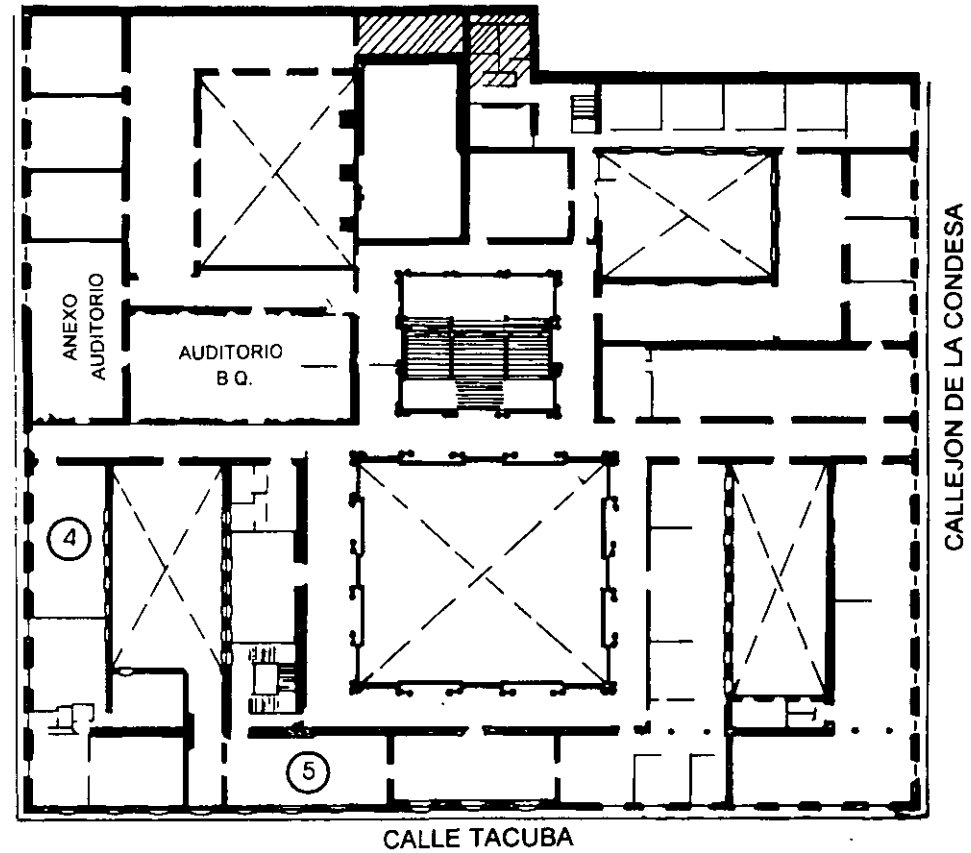
**Atentamente**

**División de Educación Continua.**

# PALACIO DE MINERIA

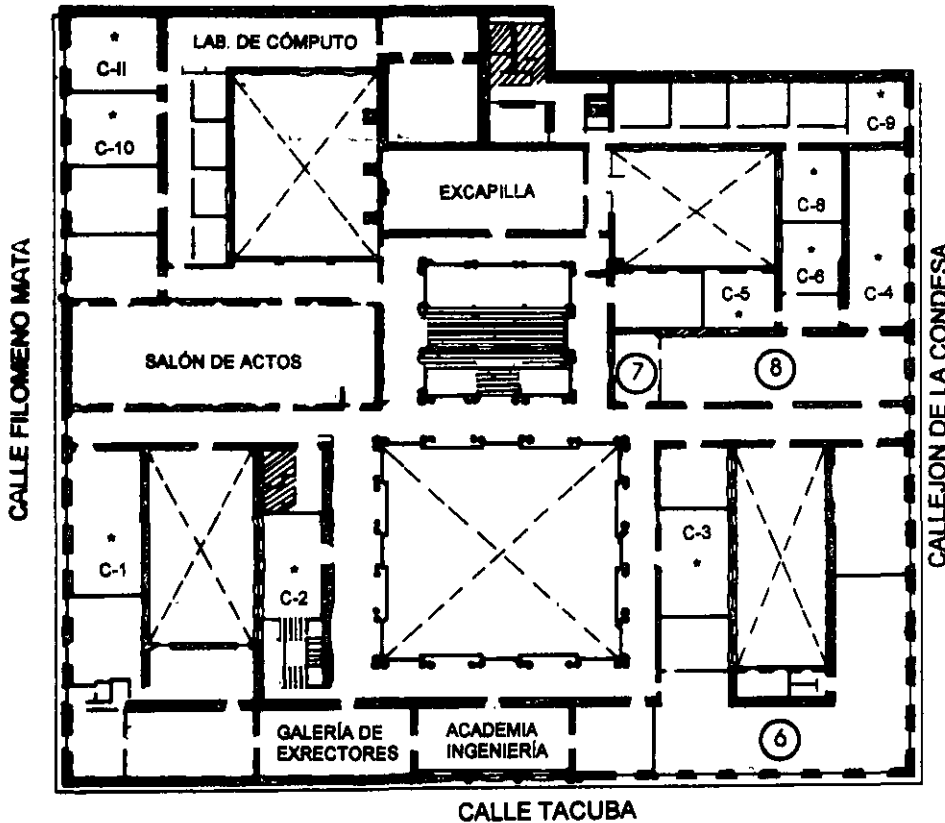


**PLANTA BAJA**



**MEZZANINNE**

# PALACIO DE MINERÍA



**1er. PISO**

## GUÍA DE LOCALIZACIÓN

1. ACCESO
2. BIBLIOTECA HISTÓRICA
3. LIBRERÍA UNAM
4. CENTRO DE INFORMACIÓN Y DOCUMENTACIÓN "ING. BRUNO MASCANZONI"
5. PROGRAMA DE APOYO A LA TITULACIÓN
6. OFICINAS GENERALES
7. ENTREGA DE MATERIAL Y CONTROL DE ASISTENCIA
8. SALA DE DESCANSO

SANITARIOS

\* AULAS



DIVISIÓN DE EDUCACIÓN CONTINUA  
FACULTAD DE INGENIERÍA U.N.A.M.  
CURSOS ABIERTOS

DIVISIÓN DE EDUCACIÓN CONTINUA









FACULTAD DE INGENIERÍA UNAM  
DIVISIÓN DE EDUCACIÓN CONTINUA

**CURSOS INSTITUCIONALES**

***BIOESTADÍSTICA:  
MÉTODOS Y  
APLICACIONES***

CI 025

***APUNTES GENERALES***

ING. JOSÉ ARTURO REYNA GALINDO  
PALACIO DE MINERÍA  
Enero del 2003

# Bioestadística: Métodos y Aplicaciones

## Índice

### 1 Conceptos previos

- 1.2 Introducción
- 1.3 ¿Que es la estadística?
- 1.5 Elementos, Población, Caracteres
- 1.7 Organización de los datos
- 1.9 Representaciones Gráficas
- 1.11 Problemas

### 2 Medidas descriptivas

- 2.1 Introducción
- 2.3 Estadísticos de tendencia central
- 2.5 Estadísticos de posición
- 2.7 Medidas de variabilidad o dispersión
- 2.9 Asimetría y apuntamiento
- 2.11 Problemas

### 3 Variables bidimensionales

- 3.2 introducción
- 3.4 Tablas de doble entrada
- 3.6 Dependencia funcional e independencia
- 3.8 Medias y varianzas marginales y condicionadas
- 3.10 Covarianza y coeficiente de correlación
- 3.12 Regresión
- 3.14 Problemas

### 4. Cálculo de probabilidades y variables aleatorias

- 4.2 introducción
- 4.4 Experimentos y sucesos aleatorios
- 4.6 Operaciones básicas con sucesos aleatorios
- 4.8 Experimentos aleatorios y probabilidad
- 4.10 Probabilidad condicionada e independencia de sucesos
- 4.12 Ciertos teoremas fundamentales del cálculo de probabilidades
- 4.14 Tests diagnósticos
- 4.16 Problemas

### 5 Variables aleatorias

- 5.2 Introducción
- 5.4 Variables aleatorias discretas
- 5.6 Variables aleatorias continuas
- <>

## 6. Principales leyes de distribución de variables aleatorias

- 6.2 Introducción
- 6.4 Distribuciones discretas
- 6.6 Reproductividad de familias de v. a.
- 6.8 Distribuciones continuas
- 6.10 Problemas

## 7. Introducción a la inferencia

- 7.2 Introducción
- 7.4 Técnicas de muestreo sobre una población
- 7.6 Propiedades deseables de un estimador

## 8. Estimación confidencial

- 8.2 Introducción
- 8.4 Intervalos de confianza para la distribución normal
- 8.6 Intervalos de confianza para variables dicotómicas
- 8.8 Problemas

## 9. Contrastes de hipótesis

- 9.2 Introducción
- 9.4 Contrastes paramétricos en una población normal
- 9.6 Contrastes de una proporción
- 9.8 Contrastes para la diferencia de medias apareadas
- 9.10 Contrastes de dos distribuciones normales independientes
- 9.12 Contrastes sobre la diferencia de proporciones
- 9.14 Problemas

## 10. Contrastes basados en el estadístico Ji-Cuadrado

- 10.2 Introducción
- 10.4 El estadístico  $\chi^2$  y su distribución
- 10.6 Contraste de bondad de ajuste para distribuciones
- 10.8 Problemas

## 11. Análisis de la varianza

- 11.2 Introducción
- 11.4 ANOVA con un factor
- 11.6 ANOVA de varios factores
- 11.8 Consideraciones sobre las hipótesis subyacentes en el modelo factorial

## 12. Contrastes no paramétricos

- 12.2 Introducción
- 12.4 Aleatoriedad de una muestra Test de rachas
- 12.6 Normalidad de una muestra Test de D'Agostino



- [12.8 Equidistribución de dos poblaciones](#)
- [12.10 Contraste de Wilcoxon para muestras apareadas](#)
- [12.12 Contraste de Kruskal-Wallis](#)
- [12.14 Problemas](#)

- [Bibliografía](#)
  - [About this document ...](#)
- 

© Este texto es la versión electrónica del manual de la Universidad de Málaga:  
*Bioestadística: Métodos y Aplicaciones*  
U.D. Bioestadística, Facultad de Medicina, Universidad de Málaga.  
ISBN. 847496-653-1

## 1. Conceptos previos

### 1.2 Introducción

Iniciamos este capítulo con la definición de algunos conceptos elementales y básicos, y sin embargo pilares, para una comprensión intuitiva y real de lo que es la Bioestadística. Pretendemos introducir al estudiante en los primeros pasos sobre el uso y manejos de datos numéricos: distinguir y clasificar las características en estudio, enseñarle a organizar y tabular las medidas obtenidas mediante la construcción de tablas de frecuencia y por último los métodos para elaborar una imagen que sea capaz de mostrar gráficamente unos resultados.

El aserto "una imagen vale más que mil palabras" se puede aplicar al ámbito de la estadística descriptiva diciendo que "un gráfico bien elaborado vale más que mil tablas de frecuencias". Cada vez es más habitual el uso de gráficos o imágenes para representar la información obtenida. No obstante, debemos ser prudente al confeccionar o interpretar gráficos, puesto que una misma información se puede representar de formas muy diversas, y no todas ellas son pertinentes, correctas o válidas. Nuestro objetivo, en este capítulo, consiste en establecer los criterios y normas mínimas que deben verificarse para construir y presentar adecuadamente los gráficos en el ámbito de la estadística descriptiva.

### 1.3 ¿Qué es la estadística?

Cuando coloquialmente se habla de estadística, se suele pensar en una relación de datos numéricos presentada de forma ordenada y sistemática. Esta idea es la consecuencia del concepto popular que existe sobre el término y que cada vez está más extendido debido a la influencia de nuestro entorno, ya que hoy día es casi imposible que cualquier medio de difusión, periódico, radio, televisión, etc. no nos aborde diariamente con cualquier tipo de información estadística sobre accidentes de tráfico, índices de crecimiento de población, turismo, tendencias políticas, etc.

Sólo cuando nos adentramos en un mundo más específico como es el campo de la investigación de las Ciencias Sociales: Medicina, Biología, Psicología, ... empezamos a percibir que la Estadística no sólo es algo más, sino que se convierte en la única herramienta que, hoy por hoy, permite dar luz y obtener resultados, y por tanto beneficios, en cualquier tipo de estudio, cuyos movimientos y relaciones, por su variabilidad intrínseca, no puedan ser abordadas desde la perspectiva de las leyes deterministas. Podríamos, desde un punto de vista más amplio, definir la estadística como la ciencia que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre.

---

La **Estadística** se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar los *datos*, siempre y cuando la variabilidad e *incertidumbre* sea una causa intrínseca de los mismos; así como de realizar *inferencias* a partir de ellos, con la finalidad de ayudar a la toma de *decisiones* y en su caso formular *predicciones*.

---

Podríamos por tanto clasificar la Estadística en descriptiva, cuando los resultados del análisis no pretenden ir más allá del conjunto de datos, e inferencial cuando el objetivo del estudio es derivar las conclusiones obtenidas a un conjunto de datos más amplio.

---

**Estadística descriptiva:** Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.

**Estadística inferencial:** Apoyándose en el cálculo de probabilidades y a partir de datos muestrales, efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

## 1.5 Elementos. Población. Caracteres

Establecemos a continuación algunas definiciones de conceptos básicos y fundamentales básicas como son: elemento, población, muestra, caracteres, variables, etc., a las cuales haremos referencia continuamente a lo largo del texto

---

**Individuos o elementos:** personas u objetos que contienen cierta información que se desea estudiar.

**Población:** conjunto de individuos o elementos que cumplen ciertas propiedades comunes.

**Muestra:** subconjunto representativo de una población.

**Parámetro:** función definida sobre los valores numéricos de características medibles de una población.

**Estadístico:** función definida sobre los valores numéricos de una muestra.

---

En relación al tamaño de la población, ésta puede ser:

- **Finita**, como es el caso del número de personas que llegan al servicio de urgencia de un hospital en un día;
- **Infinita**, si por ejemplo estudiamos el mecanismo aleatorio que describe la secuencia de caras y cruces obtenida en el lanzamiento repetido de una moneda al aire.

### 1.5.0.1 Ejemplo

Consideremos la población formada por todos los estudiantes de la Universidad de Málaga (finita). La altura media de todos los estudiantes es el parámetro  $\mu$ . El conjunto formado por los alumnos de la Facultad de Medicina es una muestra de dicha población y la altura media de esta muestra,  $\bar{x}$ , es un estadístico.

---

**Caracteres:** propiedades, rasgos o cualidades de los elementos de la población. Estos caracteres pueden dividirse en cualitativos y cuantitativos.

**Modalidades:** diferentes situaciones posibles de un carácter. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes --cada elemento posee una y sólo una de las modalidades posibles.

**Clases:** conjunto de una o más modalidades en el que se verifica que cada modalidad pertenece a una y sólo una de las clases.

## 1.7 Organización de los datos

### 1.7.2 Variables estadísticas

Cuando hablemos de **variable** haremos referencia a un símbolo (X, Y, A, B, ...) que puede tomar cualquier **modalidad** (valor) de un conjunto determinado, que llamaremos **dominio de la variable** o **rango**. En función del tipo de dominio, las variables las clasificamos del siguiente modo:

**Variables cualitativas,**

cuando las modalidades posibles son de tipo nominal. Por ejemplo, una variable de color

$$A \in \{ \text{"rojo"}, \text{"azul"}, \text{"verde"} \}.$$

#### **Variables cuasicuantitativas**

son las que, aunque sus modalidades son de tipo nominal, es posible establecer un orden entre ellas. Por ejemplo, si estudiamos la llegada a la meta de un corredor en una competición de 20 participantes, su clasificación  $C$  es tal que

$$C \in \{1^\circ, 2^\circ, 3^\circ, \dots, 20^\circ\}.$$

Otro ejemplo de variable cuasicuantitativa es el nivel de dolor,  $D$ , que sufre un paciente ante un tratamiento médico:

$$D \in \{ \text{"inexistente"}, \text{"poco intenso"}, \text{"moderado"}, \text{"fuerte"} \}.$$

#### **Variables cuantitativas**

son las que tienen por modalidades cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos grupos:

##### **Discretas,**

cuando no admiten siempre una modalidad intermedia entre dos cualesquiera de sus modalidades. Un ejemplo es el número de caras  $X$ , obtenido en el lanzamiento repetido de una moneda. Es obvio que cada valor de la variable es un número natural

$$X \in \mathbf{N}.$$

##### **Continuas,**

cuando admiten una modalidad intermedia entre dos cualesquiera de sus modalidades, v.g. el peso  $X$  de un niño al nacer. En este caso los valores de las variables son números reales. es decir

$$X \in \mathbf{R}.$$

Ocurre a veces que una variable cuantitativa continua por naturaleza, aparece como discreta. Este es el caso en que hay limitaciones en lo que concierne a la precisión del aparato de medida de esa variable, v.g. si medimos la altura en metros de personas con una regla que ofrece dos decimales de precisión, podemos obtener

$$X \in \{ \dots, 1.50, 1.51, 1.52, 1.53, \dots \}.$$

En realidad lo que ocurre es que con cada una de esas mediciones expresamos que el verdadero valor de la misma se encuentra en un intervalo de radio  $5 \cdot 10^{-3}$ . Por tanto cada una de las observaciones de  $X$  representa más bien un intervalo que un valor concreto.

Tal como hemos citado anteriormente, las modalidades son las diferentes situaciones posibles que puede presentar la variable. A veces éstas son muy numerosas (v.g. cuando

una variable es continua) y conviene reducir su número, agrupándolas en una cantidad inferior de **clases**. Estas clases deben ser construidas, tal como hemos citado anteriormente, de modo que sean *exhaustivas e incompatibles*, es decir, cada modalidad debe pertenecer a una y sólo una de las clases.

---

**Variable cualitativa:** Aquella cuyas modalidades son de tipo nominal.

**Variable cuasicuantitativa:** Modalidades de tipo nominal, en las que existe un orden.

**Variable cuantitativa discreta:** Sus modalidades son valores enteros.

**Variable cuantitativa continua:** Sus modalidades son valores reales.

### 1.7.2 Variables estadísticas

Cuando hablemos de **variable** haremos referencia a un símbolo (X.Y.A.B....) que puede tomar cualquier **modalidad** (valor) de un conjunto determinado, que llamaremos **dominio de la variable** o **rango**. En función del tipo de dominio, las variables las clasificamos del siguiente modo:

#### Variables cualitativas,

cuando las modalidades posibles son de tipo nominal. Por ejemplo, una variable de color

$$A \in \{ \text{"rojo"}, \text{"azul"}, \text{"verde"} \}.$$

#### Variables cuasicuantitativas

son las que, aunque sus modalidades son de tipo nominal, es posible establecer un orden entre ellas. Por ejemplo, si estudiamos la llegada a la meta de un corredor en una competición de 20 participantes, su clasificación  $C$  es tal que

$$C \in \{ 1^\circ, 2^\circ, 3^\circ, \dots, 20^\circ \}.$$

Otro ejemplo de variable cuasicuantitativa es el nivel de dolor.  $D$ , que sufre un paciente ante un tratamiento médico:

$$D \in \{ \text{"inexistente"}, \text{"poco intenso"}, \text{"moderado"}, \text{"fuerte"} \}.$$

#### Variables cuantitativas

son las que tienen por modalidades cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos grupos:

##### Discretas,

cuando no admiten siempre una modalidad intermedia entre dos cualesquiera de sus modalidades. Un ejemplo es el número de caras  $X$ , obtenido en el lanzamiento repetido de una moneda. Es obvio que cada valor de la variable es un número natural

$$X \in \mathbb{N}.$$

**Continuas,**

cuando admiten una modalidad intermedia entre dos cualesquiera de sus modalidades, v.g. el peso  $X$  de un niño al nacer. En este caso los valores de las variables son números reales. es decir

$$X \in \mathbb{R}.$$

Ocurre a veces que una variable cuantitativa continua por naturaleza, aparece como discreta. Este es el caso en que hay limitaciones en lo que concierne a la precisión del aparato de medida de esa variable, v.g. si medimos la altura en metros de personas con una regla que ofrece dos decimales de precisión, podemos obtener

$$X \in \{ \dots, 1.50, 1.51, 1.52, 1.53, \dots \}.$$

En realidad lo que ocurre es que con cada una de esas mediciones expresamos que el verdadero valor de la misma se encuentra en un intervalo de radio  $5 \cdot 10^{-3}$ . Por tanto cada una de las observaciones de  $X$  representa más bien un intervalo que un valor concreto.

Tal como hemos citado anteriormente, las modalidades son las diferentes situaciones posibles que puede presentar la variable. A veces éstas son muy numerosas (v.g. cuando una variable es continua) y conviene reducir su número, agrupándolas en una cantidad inferior de **clases**. Estas clases deben ser construidas, tal como hemos citado anteriormente, de modo que sean *exhaustivas e incompatibles*, es decir, cada modalidad debe pertenecer a una y sólo una de las clases.

---

**Variable cualitativa:** Aquella cuyas modalidades son de tipo nominal.

**Variable cuasicuantitativa:** Modalidades de tipo nominal, en las que existe un orden.

**Variable cuantitativa discreta:** Sus modalidades son valores enteros.

**Variable cuantitativa continua:** Sus modalidades son valores reales.

### 1.7.4 Tablas estadísticas

Consideremos una población estadística de  $n$  individuos, descrita según un carácter o variable  $C$  cuyas modalidades han sido agrupadas en un número  $k$  de clases, que denotamos

mediante  $c_1, c_2, \dots, c_k$ . Para cada una de las clases  $c_i$ ,  $i = 1, \dots, k$ , introducimos las siguientes magnitudes:

### Frecuencia absoluta

de la clase  $c_i$ , es el número  $n_i$ , de observaciones que presentan una modalidad perteneciente a esa clase.

### Frecuencia relativa

de la clase  $c_i$ , es el cociente  $f_i$ , entre las frecuencias absolutas de dicha clase y el número total de observaciones, es decir

$$f_i = \frac{n_i}{n}$$

Obsérvese que  $f_i$  es el *tanto por uno* de observaciones que están en la clase  $c_i$ .

Multiplicado por **100%** representa el porcentaje de la población que comprende esa clase.

### Frecuencia absoluta acumulada

$N_i$ , se calcula sobre variables cuantitativas o cuasicuantitativas, y es el número de elementos de la población cuya modalidad es inferior o equivalente a la modalidad  $c_i$ :

$$N_i = n_1 + n_2 + \dots + n_i = \sum_{j=1}^i n_j$$

### Frecuencia relativa acumulada

$F_i$ , se calcula sobre variables cuantitativas o cuasicuantitativas, siendo el tanto por uno de los elementos de la población que están en alguna de las clases y que presentan una modalidad inferior o igual a la  $c_i$ , es decir,

$$F_i = \frac{N_i}{n} = \frac{n_1 + \dots + n_i}{n} = f_1 + \dots + f_i = \sum_{j=1}^i f_j$$

Como todas las modalidades son exhaustivas e incompatibles ha de ocurrir que

$$\sum_{i=1}^k n_i = n_1 + n_2 + \dots + n_k = n$$

o lo que es lo mismo,

$$\sum_{i=1}^k f_i = \sum_{i=1}^k \frac{n_i}{n} = \frac{\sum_{i=1}^k n_i}{n} = \frac{n}{n} = 1.$$

**Frecuencia absoluta ( $n_i$ ):** Número de elementos que presentan la clase  $x_i$ .

Frecuencia relativa:  $f_i = n_i/N$

$$N_i = \sum_{j=1}^i n_j$$

Frecuencia absoluta acumulada:

$$F_i = N_i/N = \sum_{j=1}^i f_j$$

Frecuencia relativa acumulada:

Llamaremos **distribución de frecuencias** al conjunto de clases junto a las frecuencias correspondientes a cada una de ellas. Una **tabla estadística** sirve para presentar de forma ordenada las distribuciones de frecuencias. Su forma general es la siguiente:

Modali.	Frec. Abs.	Frec. Rel.	Frec. Abs. Acumu.	Frec. Rel. Acumu.
$C$	$n_i$	$f_i$	$N_i$	$F_i$
$c_1$	$n_1$	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = \frac{N_1}{n} = f_1$
...	...	...	...	...
$c_j$	$n_j$	$f_j = \frac{n_j}{n}$	$N_j = n_1 + \dots + n_j$	$F_j = \frac{N_j}{n} = f_1 + \dots + f_j$
...	...	...	...	...
$c_k$	$n_k$	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
	$n$	1		

### 1.7.4.1 Ejemplo

Calcular los datos que faltan en la siguiente tabla:

$l_{i-1} -- l_i$	$n_i$	$f_i$	$N_i$
0 -- 10	60	$f_1$	60
10 -- 20	$n_2$	0,4	$N_2$
20 -- 30	30	$f_3$	170
30 -- 100	$n_4$	0,1	$N_4$
100 -- 200	$n_5$	$f_5$	200
	$n$		

**Solución:**



Sabemos que la última frecuencia acumulada es igual al total de observaciones. luego  $n=200$ .

Como  $N_3=170$  y  $n_3=30$ , entonces

$$N_2=N_3-n_3=170-30=140.$$

Además al ser  $n_1=60$ , tenemos que

$$n_2=N_2-n_1=140-60=80.$$

Por otro lado podemos calcular  $n_4$  teniendo en cuenta que conocemos la frecuencia relativa correspondiente:

$$f_4 = \frac{n_4}{n} \implies n_4 = f_4 \cdot n = 0,1 \times 200 = 20$$

Así:

$$N_4=n_4+N_3=20+170=190.$$

Este último cálculo nos permite obtener

$$n_5=N_5-N_4=200-190=10.$$

Al haber calculado todas las frecuencias absolutas. es inmediato obtener las relativas:

$$\begin{aligned} f_1 &= \frac{n_1}{n} = \frac{60}{200} = 0,3 \\ f_2 &= \frac{n_2}{n} = \frac{80}{200} = 0,4 \\ f_3 &= \frac{n_3}{n} = \frac{30}{200} = 0,15 \\ f_4 &= \frac{n_4}{n} = \frac{20}{200} = 0,1 \\ f_5 &= \frac{n_5}{n} = \frac{10}{200} = 0,05 \end{aligned}$$

Escribimos entonces la tabla completa:

$l_{i-1} -- l_i$	$n_i$	$f_i$	$N_i$
0 -- 10	60	0,3	60
10 -- 20	80	0,4	140
20 -- 30	30	0,15	170
30 -- 100	20	0,1	190
100 -- 200	10	0,05	200
	200		

### 1.7.4.2 Elección de las clases

En cuanto a la elección de las clases, deben seguirse los siguientes criterios en función del tipo de variable que estudiemos:

- Cuando se trate de variables cualitativas o cuasicuantitativas, las clases  $c_i$  serán de tipo nominal;
- En el caso de variables cuantitativas, existen dos posibilidades:
  - Si la variable es discreta, las clases serán valores numéricos  $x_1, \dots, x_k$  ;
  - Si la variable es continua las clases vendrán definidas mediante lo que denominamos **intervalos**. En este caso, las modalidades que contiene una clase son todos los valores numéricos posibles contenidos en el intervalo, el cual viene normalmente definido de la forma

$$[l_{i-1}, l_i) = \{x : l_{i-1} \leq x < l_i\}$$

o bien

$$(l_{i-1}, l_i] = \{x : l_{i-1} < x \leq l_i\}$$

En estos casos llamaremos **amplitud del intervalo** a las cantidades

$$a_i = l_i - l_{i-1}$$

y **marca de clase**  $c_i$ , a un punto representativo del intervalo. Si éste es acotado, tomamos como marca de clase al punto más representativo, es decir al punto medio del intervalo,

$$c_i = \frac{l_i + l_{i-1}}{2}.$$

La marca de clase no es más que una forma abreviada de representar un intervalo mediante uno de sus puntos. Por ello hemos tomado como representante, el punto medio del mismo. Esto está plenamente justificado si recordamos que cuando se mide una variable continua como el peso, la cantidad con cierto número de decimales que expresa esta medición, no es el valor exacto de la variable, sino una medida que contiene cierto margen de error, y por tanto representa a todo un intervalo del cual ella es el centro.

En el caso de variables continuas, la forma de la tabla estadística es la siguiente:

Interv.	M. clase	Frec. Abs.	Frec. Rel.	Frec. Abs. Acum.	Frec. Rel. Acum.
	$C$	$n_i$	$f_i$	$N_i$	$F_i$
$l_0 -- l_1$	$c_1$	$n_1$	$f_1 = \frac{n_1}{n}$	$N_1 = n_1$	$F_1 = f_1$
...	...	...	...	...	...
$l_{j-1} -- l_j$	$c_j$	$n_j$	$f_j = \frac{n_j}{n}$	$N_j = N_{j-1} + n_j$	$F_j = F_{j-1} + f_j$
...	...	...	...	...	...
$l_{k-1} -- l_k$	$c_k$	$n_k$	$f_k = \frac{n_k}{n}$	$N_k = n$	$F_k = 1$
		$n$	1		

### 1.7.4.3 Elección de intervalos para variables continuas

A la hora de seleccionar los intervalos para las variables continuas, se plantean varios problemas como son el número de intervalos a elegir y sus tamaños respectivos. La notación más común que usaremos para un intervalo sea

$$l_{j-1} - l_j \stackrel{\text{def}}{=} (l_{j-1}, l_j].$$

El primer intervalo,  $l_0 - l_1$ , podemos cerrarlo en el extremo inferior para no excluir la observación más pequeña,  $l_0$

$$l_0 - l_1 \stackrel{\text{def}}{=} [l_0, l_1).$$

Éste es un convenio que tomaremos en las páginas que siguen. El considerar los intervalos por el lado izquierdo y abrirlos por el derecho no cambia de modo significativo nada de lo que expondremos.

El número de intervalos,  $k$ , a utilizar no está determinado de forma fija y por tanto tomaremos un  $k$  que nos permita trabajar cómodamente y ver bien la estructura de los datos; Como referencia nosotros tomaremos una de los siguientes valores aproximados:

$$N^\circ \text{ intervalos} \equiv k \approx \begin{cases} \sqrt{n} & \text{si } n \text{ no es muy grande,} \\ 1 + 3,22 \log n & \text{en otro caso.} \end{cases}$$

Por ejemplo si el número de observaciones que tenemos es  $n=100$ , un buen criterio es

$$k = \sqrt{100} = 10$$

agrupar las observaciones en  $k = 1 + 3,22 \log n \approx 20$  intervalos. Sin embargo si tenemos  $n=1.000.000$ , será mas razonable elegir  $k = \sqrt{1.000.000} = 1.000$  intervalos, que

La amplitud de cada intervalo

$$a_i = l_i - l_{i-1}$$

suele tomarse constante, considerando la observación más pequeña y más grande de la población (respectivamente  $l_0 = x_{\min}$  y  $l_k = x_{\max}$ ) para calcular la amplitud total,  $A$ , de la población

$$A = l_k - l_0$$

de forma que la amplitud de cada intervalo sea:

$$a_i = a \quad \forall i = 1, \dots, k \quad \text{donde} \quad a = \frac{A}{k}$$

Así la división en intervalos podría hacerse tomando:

$$\begin{aligned} l_0 &= x_{\min} \\ l_1 &= l_0 + a \\ &\dots \\ l_k &= x_{\max} = l_0 + k a \end{aligned}$$

#### 1.7.4.4 Observación

Podría ocurrir que la cantidad  $a$  fuese un número *muy desagradable* a la hora de escribir los intervalos (ej.  $a=10.325467$ ). En este caso, es recomendable variar simétricamente los extremos,  $l_0 < x_{\min} < x_{\max} < l_k$ , de forma que se tenga que  $a$  es un número más simple (ej.  $a=10$ ).

Recorrido:  $x_{\max} - x_{\min}$

Amplitud:  $a_i = l_i - l_{i-1}$

$$x_i = \frac{l_{i-1} + l_i}{2}$$

Marca de clase:

$$f_i' = \frac{n_i}{a_i}; \quad n_i' = f_i' \cdot n$$

Frecuencias rectificadas:

### 1.7.4.5 Ejemplo

Sobre un grupo de  $n=21$  personas se realizan las siguientes observaciones de sus pesos, medidos en kilogramos:

$X \sim x_1, x_2, \dots, x_{21}$						
58	42	51	54	40	39	49
56	58	57	59	63	58	66
70	72	71	69	70	68	64

Agrupar los datos en una tabla estadística.

**Solución:**

En primer lugar hay que observar que si denominamos  $X$  a la variable "peso de cada persona" esta es una variable de tipo cuantitativa y continua. Por tanto a la hora de ser ordenados los resultados en una tabla estadística, esto se ha de hacer agrupándolos en intervalos de longitud conveniente. Esto nos lleva a perder cierto grado de precisión. Para que la pérdida de información no sea muy relevante seguimos el criterio de utilizar  $k \approx \sqrt{n} = \sqrt{21}$

intervalos (no son demasiadas las observaciones). En este punto podemos tomar bien  $k=4$  o bien  $k=5$ . Arbitrariamente se elige una de estas dos posibilidades. Por ejemplo, vamos a tomar  $k=5$ .

$$\forall i = 1, \dots, 5$$

Lo siguiente es determinar la longitud de cada intervalo,  $a_i$ . Lo más cómodo es tomar la misma longitud en todos los intervalos.  $a_i = a$  (aunque esto no tiene por qué ser necesariamente así), donde

$$\begin{aligned}
 a &= \frac{A}{5} = \frac{33}{5} = 6,6 \\
 A &= l_5 - l_0 = 72 - 39 = 33 \\
 l_0 &= x_{\min} = 39 \\
 l_5 &= x_{\max} = 72
 \end{aligned}$$

Entonces tomaremos  $k=5$  intervalos de longitud  $a=6.6$  comenzando por  $l_0=x_{min}=39$  y terminando en  $l_5=72$ :

	Intervalos	M. clase	f.a.	f.r.	f.a.a.	f.r.a.
	$l_{i-1} -- l_i$	$c_i$	$n_i$	$f_i$	$N_i$	$F_i$
$i=1$	39 -- 45,6	42,3	3	0,1428	3	0.1428
$i=2$	45,6 -- 52,2	48,9	2	0,0952	5	0.2381
$i=3$	52,2 -- 58,8	55,5	6	0,2857	11	0.5238
$i=4$	58,8 -- 65,4	62,1	3	0,1428	14	0.6667
$i=5$	65,4 -- 72	68,7	7	0,3333	21	$\approx 1$
			21	$\approx 1$		

Otra posibilidad a la hora de construir la tabla, y que nos permite que trabajemos con cantidades más simples a la hora de construir los intervalos, es la siguiente. Como la regla para elegir  $l_0$  y  $l_5$  no es muy estricta podemos hacer la siguiente elección:

$$\begin{aligned}
 a' &= 7 \\
 A' &= a' \cdot 5 = 35 \\
 d &= A' - A = 35 - 33 = 2 \\
 l_0 &= x_{min} - \frac{d}{2} = 39 - 1 = 38 \\
 l_5 &= x_{max} + \frac{d}{2} = 72 + 1 = 73
 \end{aligned}$$

ya que así la tabla estadística no contiene decimales en la expresión de los intervalos, y el exceso  $d$ , cometido al ampliar el rango de las observaciones desde  $A$  hasta  $A'$ , se reparte del mismo modo a los lados de las observaciones menores y mayores:

	Intervalos	M. clase	f.a.	f.r.	f.a.a.	f.r.a.
	$l_{i-1} -- l_i$	$c_i$	$n_i$	$f_i$	$N_i$	$F_i$
$i=1$	38 -- 45	41,5	3	0,1428	3	0,1428
$i=2$	45 -- 52	48,5	2	0,0952	5	0,2381
$i=3$	52 -- 59	55,5	7	0,3333	12	0,5714
$i=4$	59 -- 66	62,5	3	0,1428	15	0,7143
$i=5$	66 -- 73	69,5	6	0,2857	21	$\approx 1$
			21	$\approx 1$		

## 1.9 Representaciones Gráficas

Hemos visto que la tabla estadística resume los datos que disponemos de una población. de forma que ésta se puede analizar de una manera más sistemática y resumida . Para darnos cuenta *de un sólo vistazo* de las características de la población resulta aún más esclarecedor el uso de gráficos y diagramas, cuya construcción abordamos en esta sección.

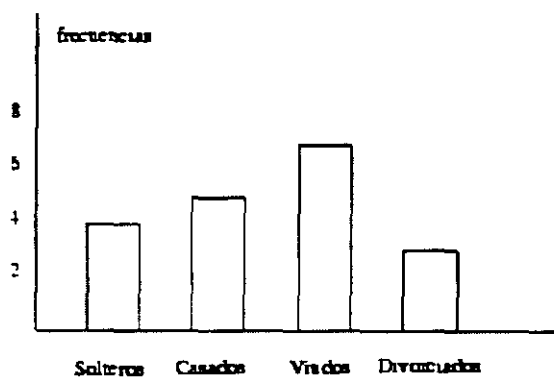
### 1.9.2 Gráficos para variables cualitativas

Los gráficos más usuales para representar variables de tipo nominal son los siguientes:

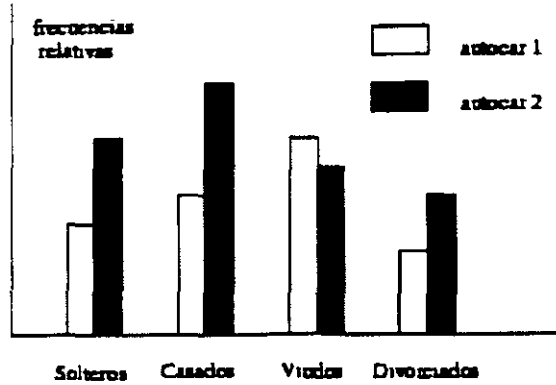
#### Diagramas de barras:

Siguiendo la figura 1.1, representamos en el eje de ordenadas las modalidades y en abscisas las frecuencias absolutas o bien, las frecuencias relativas. Si, mediante el gráfico, se intenta comparar varias poblaciones entre sí, existen otras modalidades, como las mostradas en la figura 1.2. Cuando los tamaños de las dos poblaciones son diferentes, es conveniente utilizar las frecuencias relativas, ya que en otro caso podrían resultar engañosas.

**Figura:** Diagrama de barras para una variable cualitativa.



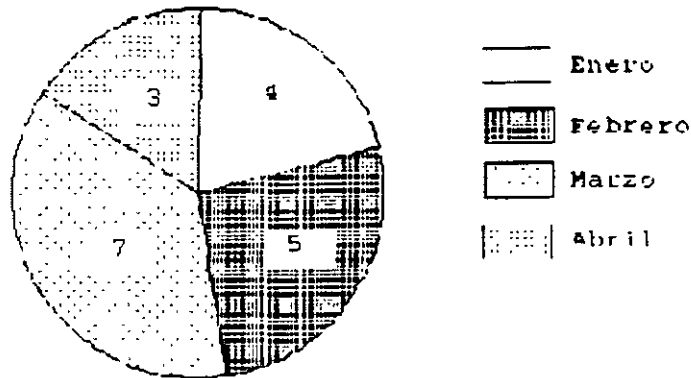
**Figura:** Diagramas de barras para comparar una variable cualitativa en diferentes poblaciones. Se ha de tener en cuenta que la altura de cada barra es *proporcional* al número de observaciones (frecuencias relativas).



### Diagramas de sectores

(también llamados *tartas*). Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa (figura 1.3).

Figura: Diagrama de sectores.



El arco de cada porción se calcula usando la *regla de tres*:

$$\begin{aligned}
 n &\rightarrow 360^\circ \\
 n_i &\rightarrow x_i = \frac{360 \cdot n_i}{n}
 \end{aligned}$$

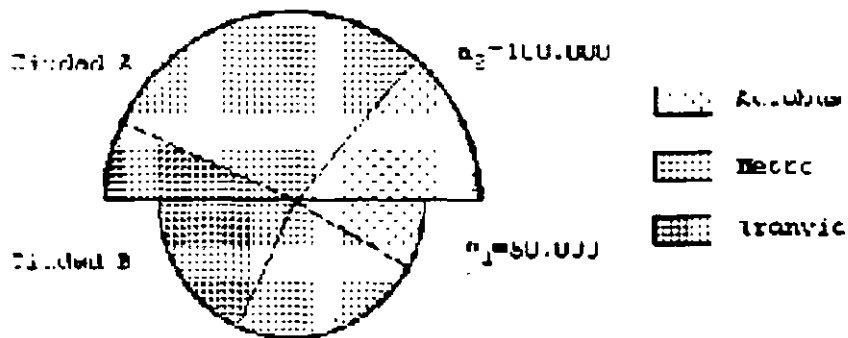
Como en la situación anterior, puede interesar comparar dos poblaciones. En este caso también es aconsejable el uso de las frecuencias relativas (porcentajes) de ambas sobre gráficos como los anteriores. Otra posibilidad es comparar las 2 poblaciones usando para cada una de ellas un diagrama semicircular, al igual que en



la figura 1.4. Sean  $n_1 \leq n_2$  los tamaños respectivos de las 2 poblaciones. La población más pequeña se representa con un semicírculo de radio  $r_1$  y la mayor con otro de radio  $r_2$ . La relación existente entre los radios, es la que se obtiene de suponer que la relación entre las áreas de las circunferencias es igual a la de los tamaños de las poblaciones respectivas, es decir:

$$\frac{r_2^2}{r_1^2} = \frac{n_2}{n_1} \iff r_2 = r_1 \cdot \sqrt{\frac{n_2}{n_1}}$$

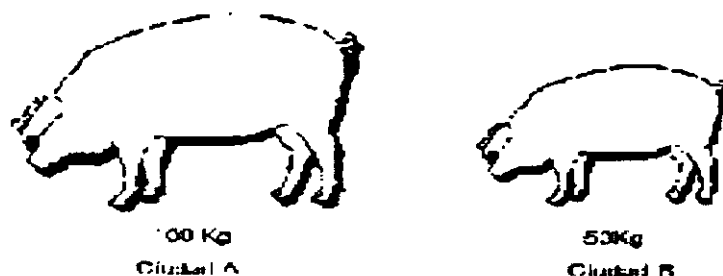
**Figura:** Diagrama de sectores para comparar dos poblaciones  
Medios de transporte preferidos



### Pictogramas

Expresan con dibujos alusivo al tema de estudio las frecuencias de las modalidades de la variable. Estos gráficos se hacen representado a diferentes escalas un mismo dibujo, como vemos en la figura 1.5.

**Figura:** Pictograma. Las áreas son proporcionales a las frecuencias.  
 UNIDADES DE COMODIDAD POR DÍA EN UN FIN DE SEMANA



El escalamiento de los dibujos debe ser tal que el *área*<sup>11</sup> de cada uno de ellos sea proporcional a la frecuencia de la modalidad que representa. Este tipo de gráficos suele usarse en los medios de comunicación, para que sean comprendidos por el público no especializado, sin que sea necesaria una explicación compleja.

### 1.9.4 Gráficos para variables cuantitativas

Para las variables cuantitativas, consideraremos dos tipos de gráficos, en función de que para realizarlos se usen las frecuencias (absolutas o relativas) o las frecuencias acumuladas:

#### Diagramas diferenciales:

Son aquellos en los que se representan frecuencias absolutas o relativas. En ellos se representa el número o porcentaje de elementos que presenta una modalidad dada.

#### Diagramas integrales:

Son aquellos en los que se representan el número de elementos que presentan una modalidad inferior o igual a una dada. Se realizan a partir de las frecuencias acumuladas, lo que da lugar a gráficos crecientes, y es obvio que este tipo de gráficos no tiene sentido para variables cualitativas.

Según hemos visto existen dos tipos de variables cuantitativas: discretas y continuas. Vemos a continuación las diferentes representaciones gráficas que pueden realizarse para cada una de ellas así como los nombres específicos que reciben.

#### 1.9.4.1 Gráficos para variables discretas

Cuando representamos una variable discreta, usamos el **diagrama de barras** cuando pretendemos hacer una gráfica diferencial. Las barras deben ser estrechas para representar el que los valores que toma la variable son discretos. El diagrama integral o acumulado tiene, por la naturaleza de la variable, forma de escalera. Un ejemplo de diagrama de barras así como su diagrama integral correspondiente están representados en la figura 1.6.

#### 1.9.4.2 Ejemplo

Se lanzan tres monedas al aire en 8 ocasiones y se contabiliza el número de caras,  $X$ , obteniéndose los siguientes resultados:

$$X \sim 2, 1, 0, 1, 3, 2, 1, 2$$

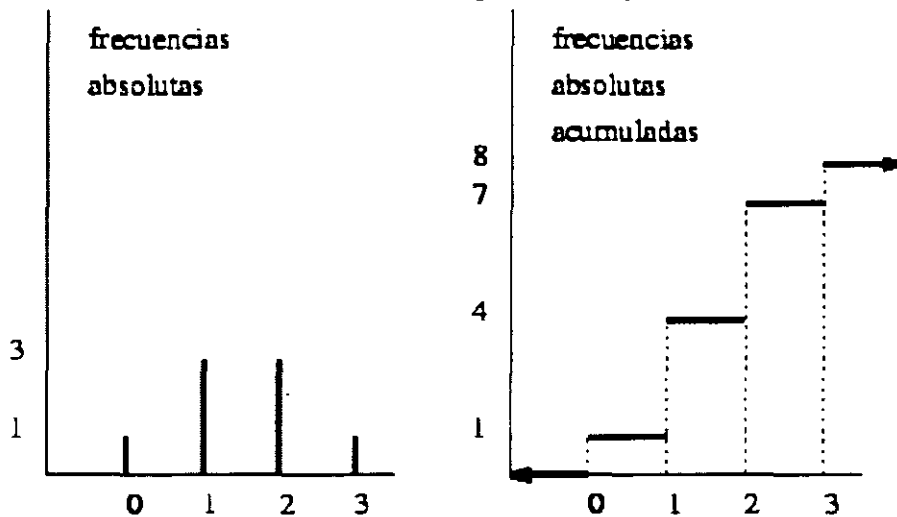
Representar gráficamente el resultado.

**Solución:** En primer lugar observamos que la variable  $X$  es cuantitativa discreta, presentando las modalidades:

$$X \in 0, 1, 2, 3$$

Ordenamos a continuación los datos en una tabla estadística, y se representa la misma en la figura 1.6.

**Figura:** Diagrama diferencial (barras) e integral para una variable discreta  
 Obsérvese que el diagrama integral (creciente) contabiliza el número de observaciones de la variable inferiores o iguales a cada punto del eje de abscisas.



$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
0	1	1/8	1	1/8
1	3	3/8	4	4/8
2	3	3/8	7	7/8
3	1	1/8	8	8/8
	$n=8$	1		

### 1.9.4.3 Ejemplo

Clasificadas 12 familias por su número de hijos se obtuvo:

Número de hijos ( $x_i$ )	1	2	3	4
Frecuencias ( $n_i$ )	1	3	5	3

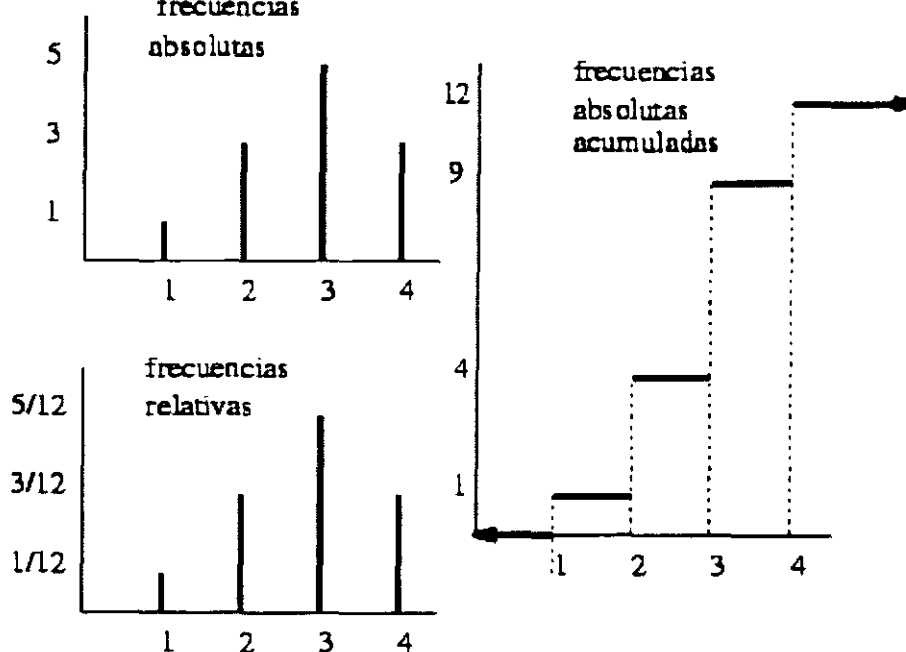
Comparar los diagramas de barras para frecuencias absolutas y relativas. Realizar el diagrama acumulativo creciente.

**Solución:** En primer lugar, escribimos la tabla de frecuencias en el modo habitual:

Variable	F. Absolutas	F. Relativas	F. Acumuladas
$x_i$	$n_i$	$f_i$	$N_i$
1	1	0,083	1
2	3	0,250	4
3	5	0,416	9
4	3	0,250	12
	12	1	

Con las columnas relativas a  $x_i$  y  $n_i$  realizamos el diagrama de barras para frecuencias absolutas, lo que se muestra en la figura 1.7. Como puede verse es idéntico (salvo un cambio de escala en el eje de ordenadas) al diagrama de barras para frecuencias relativas y que ha sido calculado usando las columnas de  $x_i$  y  $f_i$ . El diagrama escalonado (acumulado) se ha construido con la información procedente de las columnas  $x_i$  y  $N_i$ .

**Figura:** Diagramas de frecuencias para una variable discreta



#### 1.9.4.4 Gráficos para variables continuas

Cuando las variables son continuas, utilizamos como diagramas diferenciales los *histogramas* y los *polígonos de frecuencias*.

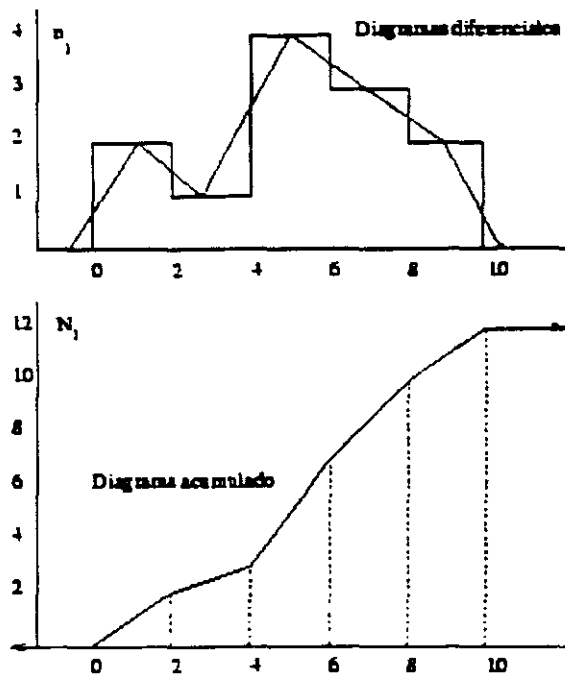
Un *histograma* se construye a partir de la tabla estadística, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y el área de los mismos.

El *polígono de frecuencias* se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase. Para representar el polígono de frecuencias en el primer y último intervalo, suponemos que adyacentes a ellos existen otros intervalos de la misma amplitud y frecuencia nula, y se unen por una línea recta los puntos del histograma que corresponden a sus marcas de clase. Obsérvese que de este modo, el polígono de frecuencias tiene en común con el histograma el que las áreas de la gráficas sobre un intervalo son idénticas. Veanse ambas gráficas diferenciales representadas en la parte superior de la figura 1.8.

El diagrama integral para una variable continua se denomina también **polígono de frecuencias acumulado**, y se obtiene como la poligonal definida en abscisas a partir de los extremos de los intervalos en los que hemos organizado la tabla de la variable, y en ordenadas por alturas que son proporcionales a las frecuencias acumuladas. Dicho de otro modo, el polígono de frecuencias absolutas es una primitiva del histograma. Véase la parte inferior de la figura 1.8, en la que se representa a modo de ilustración los diagramas correspondientes a la variable cuantitativa continua expresada en la tabla siguiente:

Intervalos	$c_i$	$n_i$	$N_i$
0 -- 2	1	2	2
2 -- 4	3	1	3
4 -- 6	5	4	7
6 -- 8	7	3	10
8 - 10	9	2	12
		12	

**Figura:** Diagramas diferenciales e integrales para una variable continua.



### 1.9.4.5 Ejemplo

La siguiente distribución se refiere a la duración en horas (completas) de un lote de 500 tubos:

Duración en horas	Número de tubos
300 -- 500	50
500 -- 700	150
700 -- 1.100	275
más de 1.100	25
	Total 500

- Representar el histograma de frecuencias relativas y el polígono de frecuencias.
- Trazar la curva de frecuencias relativas acumuladas.
- Determinar el número mínimo de tubos que tienen una duración inferior a 900 horas.

**Solución:** En primer lugar observamos que la variable en estudio es discreta (*horas completas*), pero al tener un rango tan amplio de valores resulta más conveniente agruparla en intervalos, como si de una variable continua se tratase. La consecuencia es una ligera pérdida de precisión.

El último intervalo está abierto por el límite superior. Dado que en él hay 25 observaciones puede ser conveniente cerrarlo con una amplitud "razonable". Todos los intervalos excepto

el tercero tienen una amplitud de 200 horas, luego podríamos cerrar el último intervalo en 1.300 horas<sup>1,2</sup>.

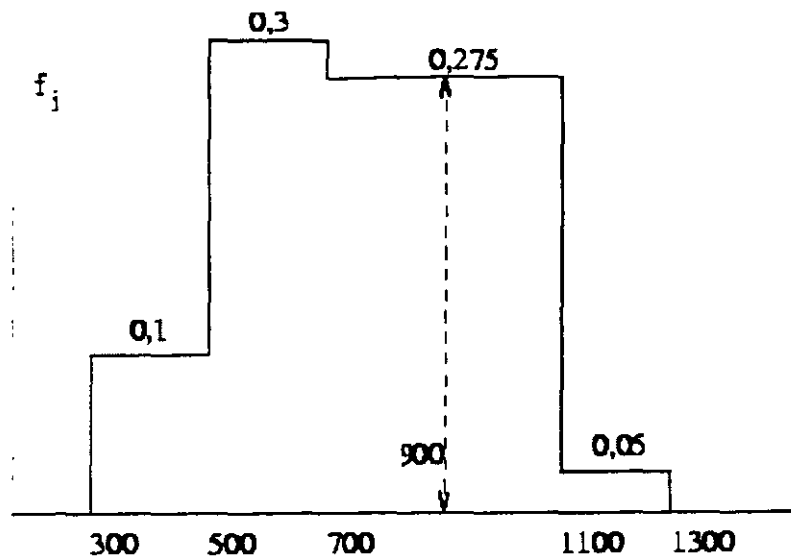
Antes de realizar el histograma conviene hacer una observación importante. El histograma representa las frecuencias de los intervalos mediante *áreas* y no mediante *alturas*. Sin embargo nos es mucho más fácil hacer representaciones gráficas teniendo en cuenta estas últimas. Si todos los intervalos tienen la misma amplitud no es necesario diferenciar entre los conceptos de área y altura, pero en este caso el tercer intervalo tiene una amplitud doble a los demás, y por tanto hay que repartir su área en un rectángulo de base doble (lo que reduce su altura a la mitad).

Así será conveniente añadir a la habitual tabla de frecuencias una columna que represente a las amplitudes,  $a_i$ , de cada intervalo, y otra de frecuencias relativas rectificadas,  $f'_i$ , para representar la altura del histograma. Los gráficos requeridos se representan en las figuras 1.9 y 1.10.

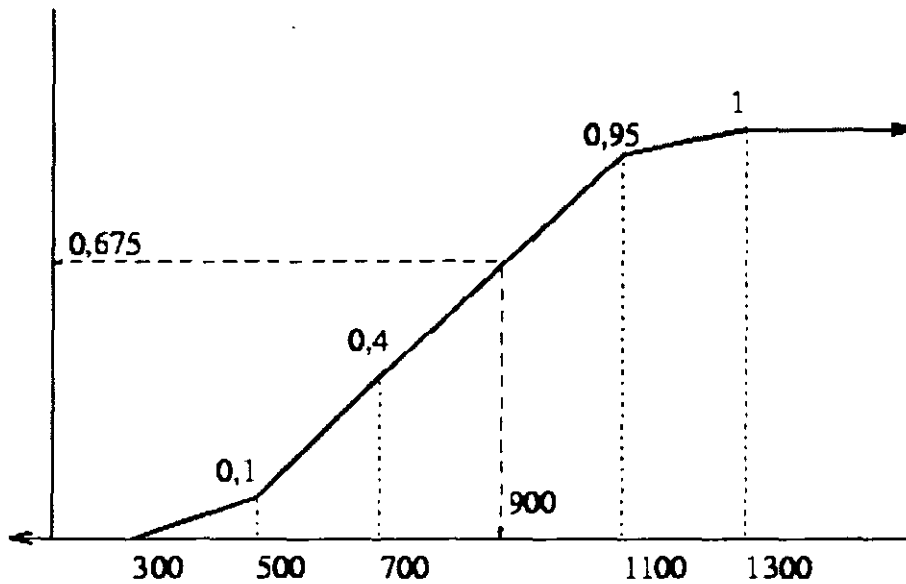
Intervalos	$a_i$	$n_i$	$f_i$	$f'_i$	$F_i$
300 -- 500	200	50	0.10	0,10	0.10
500 -- 700	200	150	0.30	0.30	0.40
700 -- 1.100	400	275	0.55	0,275	0.95
1.100 -- 1.300	200	25	0.05	0.05	1.00
		n=500			

**Figura:** Histograma. Obsérvese que la altura del histograma en cada intervalo es  $f'_i$  que coincide en todos con  $f_i$  salvo en el intervalo 700 --

1.100 en el que  $f'_i = 1/2 f_i$  ya que la amplitud de ese intervalo es doble a la de los demás.



**Figura:** Diagrama acumulativo de frecuencias relativas



Por otro lado, mirando la figura 1.9 se ve que sumando frecuencias relativas, hasta las 900 horas de duración hay

$$0,10 + 0,30 + 0,275 = 0,675 = 67,5 \% \text{ de los tubos.}$$

Esta cantidad se obtiene de modo más directo viendo a qué altura corresponde al valor 900 en el diagrama de frecuencias acumuladas (figura 1.10).



Como en total son 500 tubos. el número de tubos con una duración igual o menor que 900 horas es  $0,675 \times 500 = 337,5$ , redondeando, 338 tubos.

**Tabla:** Principales diagramas según el tipo de variable.

Tipo de variable	Diagrama
V. Cualitativa	Barras, sectores, pictogramas
V. Discreta	Diferencial (barras) Integral (en escalera)
V. Continua	Diferencial (histograma, polígono de frecuencias) Integral (diagramas acumulados)

## 1.11 Problemas

**Ejercicio 1.1.** Clasificar las siguientes variables:

1. Preferencias políticas (izquierda, derecha o centro).
2. Marcas de cerveza.
3. Velocidad en Km/h.
4. El peso en Kg.
5. Signo del zodiaco.
6. Nivel educativo (primario secundario, superior).
7. Años de estudios completados.

8. Tipo de enseñanza (privada o pública).
9. Número de empleados de una empresa.
10. La temperatura de un enfermo en grados Celsius.
11. La clase social (baja, media o alta).
12. 
$$\text{Nw/cm}^2$$
La presión de un neumático en

**Ejercicio 1..2.** Clasifique las variables que aparecen en el siguiente cuestionario.

1. ¿Cuál es su edad?
2. Estado civil:
  - (a) Soltero
  - (b) Casado
  - (c) Separado
  - (d) Divorciado
  - (e) Viudo
3. ¿Cuanto tiempo emplea para desplazarse a su trabajo?
4. Tamaño de su municipio de residencia:
  - (a) Municipio pequeño (menos de 2.000 habitantes)
  - (b) Municipio mediano (de 2.000 a 10.000 hab.)
  - (c) Municipio grande (de 10.000 a 50.000 hab.)
  - (d) Ciudad pequeña (de 50.000 a 100.000 hab.)
  - (e) Ciudad grande (más de 100.000 hab.)
5. ¿Está afiliado a la seguridad social?

### Ejercicio 1.3.

En el siguiente conjunto de datos, se proporcionan los pesos (redondeados a libras) de niños nacidos en cierto intervalo de tiempo:

4, 8, 4, 6, 8, 6, 7, 7, 7, 8, 10, 9, 7, 6, 10, 8, 5, 9, 6, 3, 7, 6, 4, 7, 6, 9, 7, 4, 7, 6, 8, 8, 9, 11, 8, 7, 10, 8, 5, 7, 7, 6, 5, 10, 8, 9, 7, 5, 6, 5.

1. Construir una distribución de frecuencia de estos pesos.
2. Encontrar las frecuencias relativas.
3. Encontrar las frecuencias acumuladas.
4. Encontrar las frecuencias relativas acumuladas.
5. Dibujar un histograma con los datos del apartado a.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?