



Escuela Nacional de
Medicina del Trabajo

CURSO DE INVESTIGACIÓN CIENTÍFICA

en Medicina del Trabajo
y Salud Laboral



Instituto Navarro de
Salud Laboral

Estadística Básica – Parte I

Notas Basadas en Manual “Bioestadística: Métodos y aplicaciones”
Facultad de Medicina. [Universidad de Málaga](http://www.unimálaga.es).

Guía de Clase

Contenido

1. CONCEPTOS PREVIOS	1
1.1 Introducción	1
1.2 ¿Qué es la estadística?	1
1.3 Elementos: población, variables	1
1.4 Organización de los datos	2
1.4.1 Variables estadísticas	2
1.4.2 Tablas estadísticas	3
1.5 Representaciones Gráficas	9
1.5.1 Gráficos para variables cualitativas	9
1.5.2 Gráficos para variables cuantitativas	10
1.6 Problemas	15
2. MEDIDAS DESCRIPTIVAS	17
2.1 Introducción	17
2.2 Estadísticos de tendencia central	18
2.2.1 La media	18
2.2.2 La mediana	21
2.2.3 La moda	24
2.2.4 Relación entre media, mediana y moda	25
2.3 Estadísticos de posición	27
2.3.1 Ejemplo	27
2.3.2 Ejemplo	28
2.3.3 Ejemplo	29
2.3.4 Ejemplo	30
2.4 Medidas de variabilidad o dispersión	32
2.4.1 Desviación media, D_m	32
2.4.2 Varianza y desviación típica	32
2.4.3 Coeficiente de variación	38
2.5 Problemas	40
3. VARIABLES BIDIMENSIONALES	43
3.1 Introducción	43
3.2 Tablas de doble entrada	44
3.3 Dependencia funcional e independencia	45

Parte I - Contenidos

3.3.1 Dependencia funcional	45
3.3.2 Independencia	46
BIBLIOGRAFÍA	1
BIBLIOGRAFÍA ADICIONAL	1

1. Conceptos previos

1.1 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 el uso y manejo 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.

1.2 ¿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, sicología, empezamos a percibir que la Estadística no sólo es algo más, sino que se convierte en una herramienta casi única 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 disciplina que estudia cómo debe emplearse la información y cómo dar una guía de acción en situaciones prácticas que entrañan incertidumbre.

1.3 Elementos: población, variables

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

- **Individuos, elementos o unidades de análisis:** personas u objetos que contienen cierta información que se desea estudiar.
- **Población:** conjunto de individuos o elementos que cumplen ciertas propiedades comunes.
- **Muestra:** subconjunto 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.

Ejemplo

Consideremos la población formada por todos los estudiantes del curso (finita). La altura media de todos los estudiantes es el parámetro μ . El conjunto formado por los alumnos hombres 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 o categorías:** 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 o categorías posibles.
- **Clases:** conjunto de una o más categorías en el que se verifica que cada unidad pertenece a una y sólo una modalidad.

1.4 Organización de los datos

1.4.1 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:

1. Variables cualitativas,

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

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

2. Variables cuasi-cuantitativas u ordinales

Son las que, aunque sus modalidades sean 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 cuasi-cuantitativa es el nivel de dolor, D, que sufre un paciente ante un tratamiento médico:

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

3. 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 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

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

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.

En resumen, estos son los tipos de variables:

- **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.

Desde el punto de vista epidemiológico, las variables cualitativas se clasifican en nominales y ordinales, y las cuantitativas en variables de intervalo y de razón. Estos cuatro niveles son importantes porque jerarquizan las posibilidades de análisis, así el nivel más alto lo permiten las variables cuantitativas de razón y el más simple las nominales. Esta clasificación, además de permitir la elección más acertada de los procedimientos estadísticos a utilizar, simplifica el planteamiento de las hipótesis, la combinación de las pruebas y la generación de modelos.

Cualitativas	Nominales Ordinales
Cuantitativas	De interval De razón

1.4.2 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**

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_j :

- **Frecuencia relativa acumulada**

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_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,

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

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

2. **Frecuencia relativa:** $f_i = n_i/N$

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

3. **Frecuencia absoluta acumulada:**

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

4. **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		

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:

$$f_1 = \frac{n_1}{n} = \frac{60}{200} = 0,3$$

$$f_3 = \frac{n_3}{n} = \frac{30}{200} = 0,15$$

$$f_5 = \frac{n_5}{n} = \frac{10}{200} = 0,05$$

Escribimos entonces la tabla completa:

$li-1 - li$	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		

1.4.2.1 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

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		

1.4.2.2 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.

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

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 agrupar las observaciones en $k = \sqrt{100} = 10$ intervalos. Sin embargo si tenemos $n=1.000.000$, será mas

razonable elegir $k = 1 + 3,22 \log n \approx 20$ intervalos, que $k = \sqrt{1.000.000} = 1.000$.

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
 $A = l_k - l_0$

Así la división en intervalos podría hacerse tomando:

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

1.4.2.3 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:

Frecuencias rectificadas: $f_i' = \frac{n_i}{a_i}$; $n_i' = f_i' \cdot n$

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 $k=5$.

Lo siguiente es determinar la longitud de cada intervalo, a_i , $\forall i = 1, \dots, 5$. 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	≈ 1
			21	≈ 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	≈ 1
			21	≈ 1		

1.5 Representaciones Gráficas

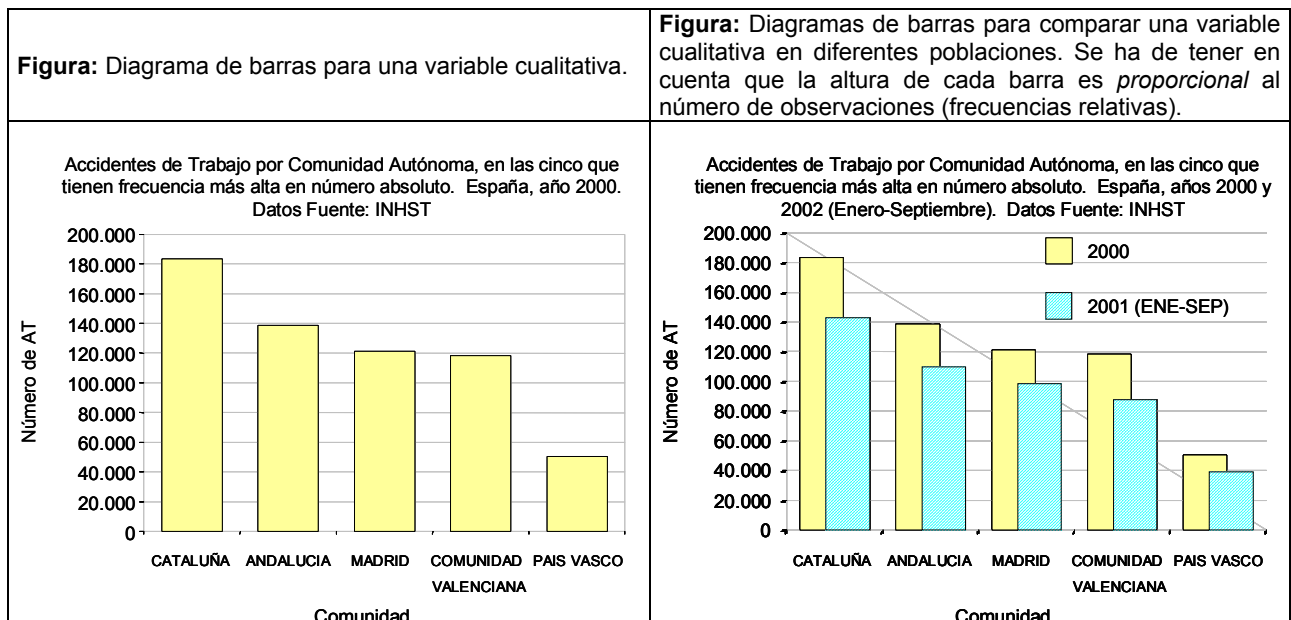
Hemos visto que la tabla estadística resume los datos que disponemos de una población, de forma que ésta se puede analizar de una manera más sistemática y resumida. Para darnos cuenta de un sólo vistazo de las características de la población resulta aún más esclarecedor el uso de gráficos y diagramas,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

1.5.1 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.

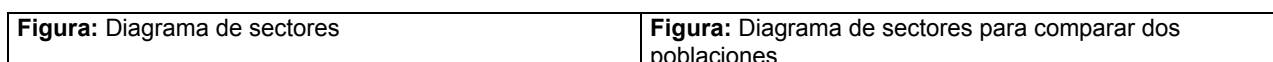


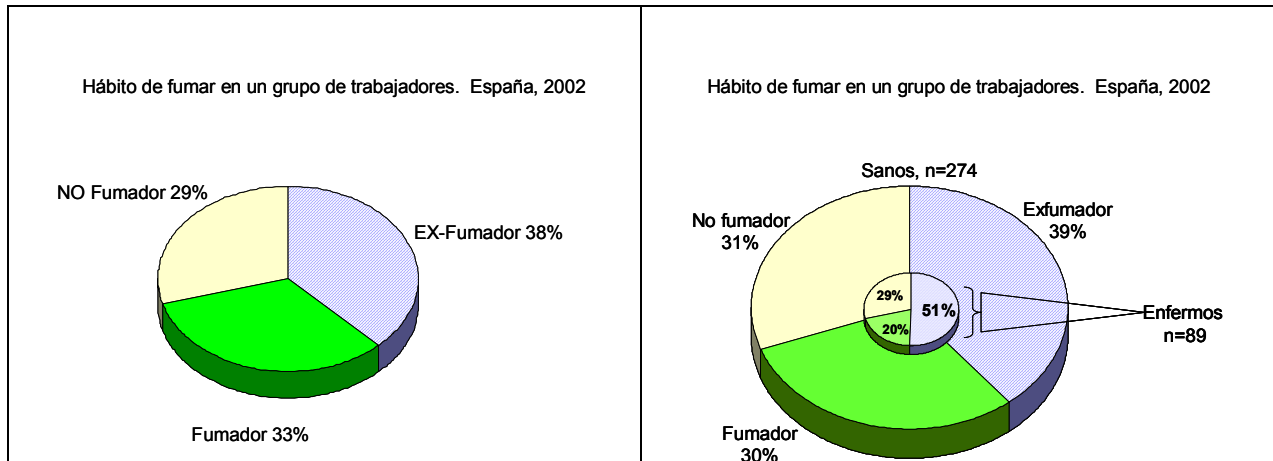
Diagramas de sectores

(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).

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

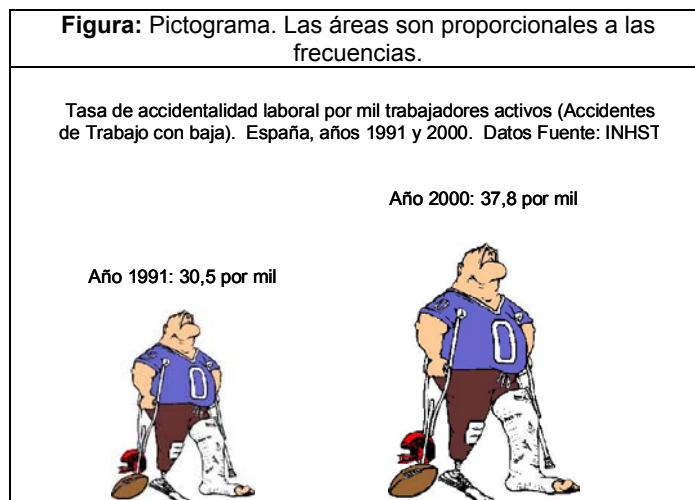
semicircular, al igual que en la figura 1.4. Sean $n_1 \leq n_2$ los tamaños respectivos de las 2 poblaciones. La población más pequeña se representa con un semicírculo de radio r_1 y la mayor con otro de radio r_2 . La relación existente entre los radios, es la que se obtiene de suponer que la relación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





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.



El escalamiento de los dibujos debe ser tal que el área 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.

1.5.2 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.

1.5.2.1 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.

Ejemplo

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

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

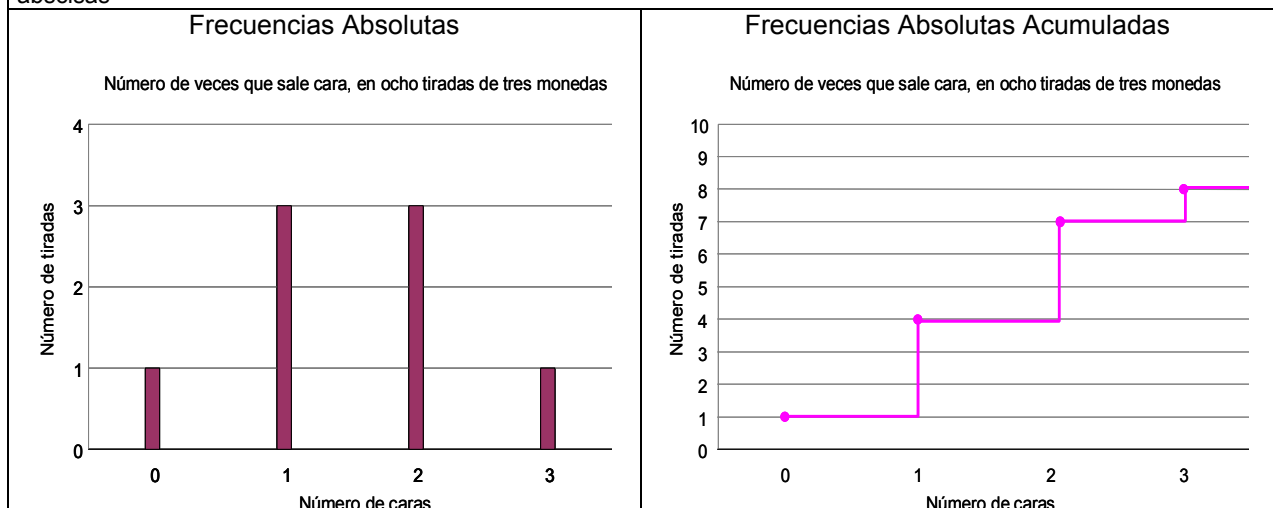
Representar gráficamente el resultado.

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		

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

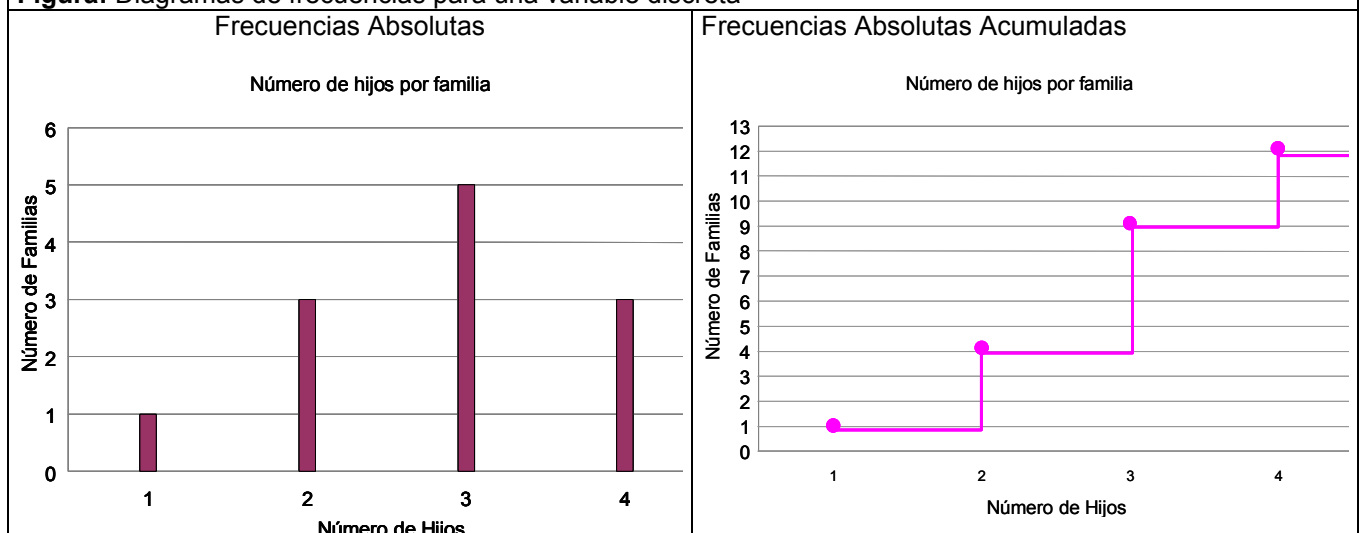
Comparar los diagramas de barras para frecuencias absolutas y relativas. Realizar el diagrama acumulativo creciente.

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



1.5.2.2 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.

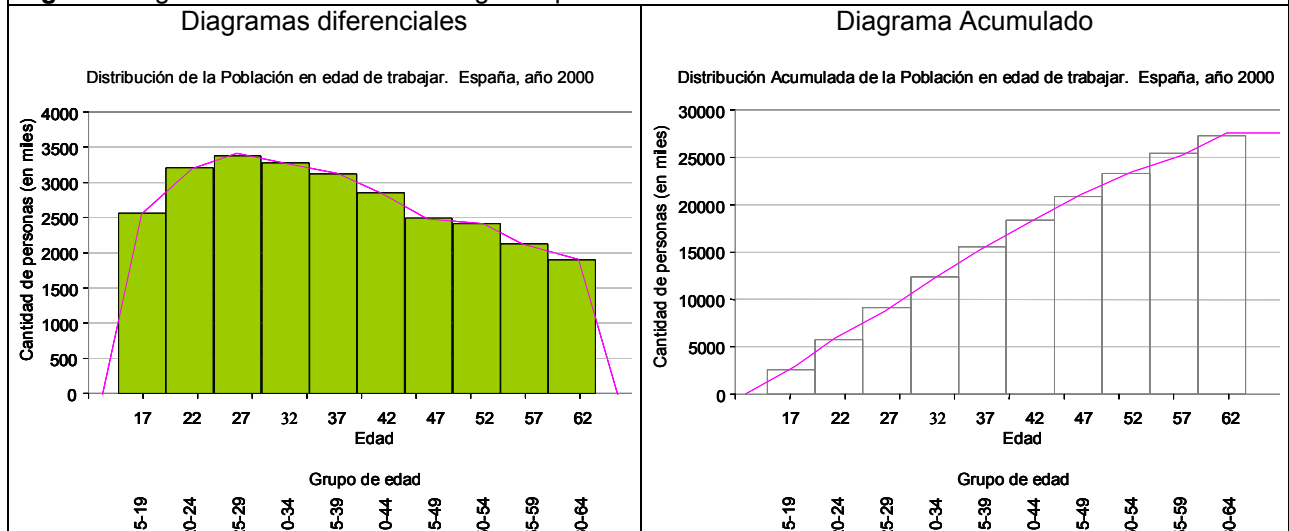
El polígono de frecuencias se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase. Para representar el polígono de frecuencias en el primer y último intervalo, suponemos 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. Ver 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 edad, de la población que puede vincularse al proceso productivo:

Edad	C_i	n_i	N_i
15-19	17	2564	2564
20-24	22	3208	5772
25-29	27	3373	9145
30-34	32	3276	12421
35-39	37	3123	15544
40-44	42	2847	18391
45-49	47	2493	20884
50-54	52	2409	23293
55-59	57	2131	25424
60-64	62	1899	27323

Figura: Diagramas diferenciales e integrales para una variable continua.



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^{1,2}.

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.

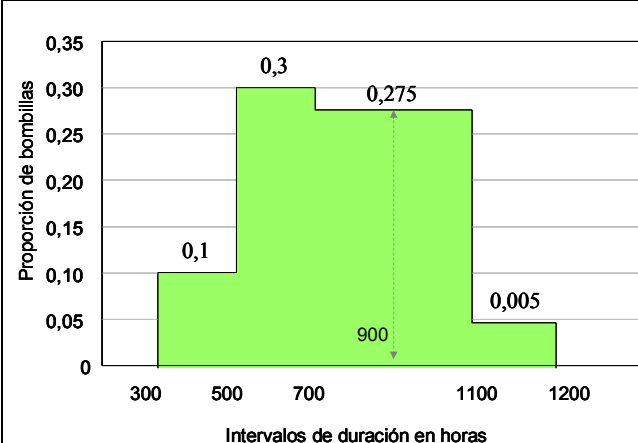
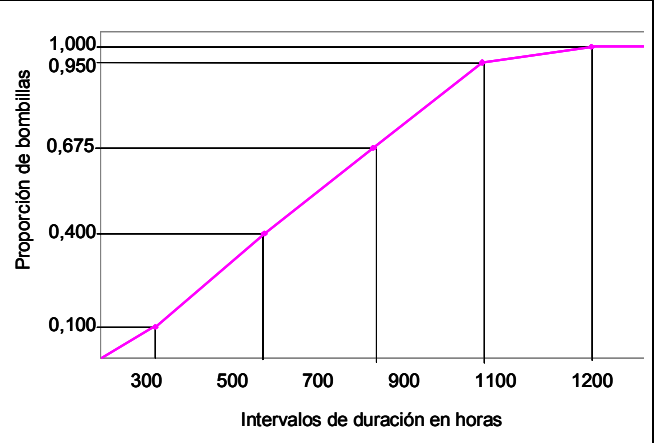


Figura: Diagrama acumulativo de frecuencias relativas



Por otro lado, mirando la figura 1.9 se ve que sumando frecuencias relativas, hasta las 900 horas de duración hay $0,10 + 0,30 + 0,275 = 0,675 = 67,5\%$ 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)

1.6 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. La presión de un neumático en Nw/cm^2

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.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?

2. Medidas descriptivas

2.1 Introducción

Los fenómenos biológicos no suelen ser constantes, por lo que será necesario que junto a una medida que indique el valor alrededor del cual se agrupan los datos, se asocie una medida que haga referencia a la variabilidad que refleje dicha fluctuación.

En este sentido pueden examinarse varias características, siendo las más comunes:

La [tendencia central](#) de los datos;

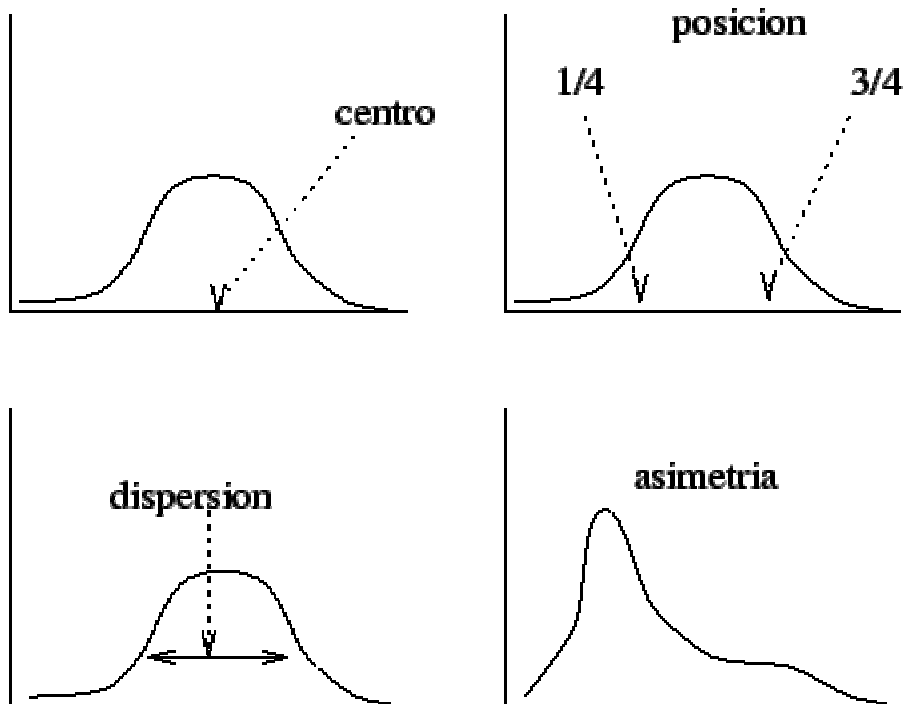
La [dispersión o variación](#) con respecto a este centro;

Los datos que ocupan ciertas [posiciones](#).

La [simetría](#) de los datos.

La [forma](#) en la que los datos se agrupan.

Figura: Medidas representativas de un conjunto de datos estadísticos



A lo largo de este capítulo, y siguiendo este orden, iremos estudiando los estadísticos que nos van a orientar sobre cada uno de estos niveles de información: valores alrededor de los cuales se agrupa la muestra, la mayor o menor fluctuación alrededor de esos valores, nos interesaremos en ciertos valores que marcan posiciones características de una distribución de frecuencias así como su simetría y su forma.

2.2 Estadísticos de tendencia central

Las tres medidas más usuales de tendencia central son:

la [media](#),
la [mediana](#),
la [moda](#).

En ciertas ocasiones estos tres estadísticos suelen coincidir, aunque generalmente no es así. Cada uno de ellos presenta [ventajas e inconvenientes](#).

2.2.1 La media

La media aritmética de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos. Es decir, si la tabla de valores de una variable X es

X	n_i	f_i
x_1	n_1	f_1
...
x_k	n_k	f_k

la media es el valor que podemos escribir de las siguientes formas equivalentes:

$$\begin{aligned} \bar{x} &= x_1 f_1 + \dots + x_k f_k \\ &= \frac{1}{n} (x_1 n_1 + \dots + x_k n_k) \\ &= \frac{1}{n} \sum_{i=1}^k x_i n_i \end{aligned}$$

Si los datos no están ordenados en una tabla, entonces

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

2.2.1.1 Proposición

La suma de las *diferencias de la variable con respecto a la media* es nula, es decir,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Demostración

Basta desarrollar el sumatorio para obtener

$$\sum_{i=1}^n (x_i - \bar{x}) = (x_1 - \bar{x}) + \dots + (x_n - \bar{x}) = (x_1 + \dots + x_n) - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

Este resultado nos indica que el error cometido al aproximar un valor cualquiera de la variable, por ejemplo x_1 , mediante el valor central \bar{x} , es compensado por los demás errores:

$$\text{Error aprox. de } x_1 \quad \equiv \quad x_1 - \bar{x} = \sum_{i=2}^n (x_i - \bar{x})$$

Si los errores se consideran con signo positivo, en este caso no pueden compensarse. Esto ocurre si tomamos como medida de error alguna de las siguientes:

$$\sum_{i=1}^n (x_i - \bar{x})^2 \geq 0 \quad \text{Error cuadrático}$$

$$\sum_{i=1}^n |x_i - \bar{x}| \geq 0 \quad \text{Error absoluto}$$

$$\max_{i=1, \dots, n} |x_i - \bar{x}| \geq 0 \quad \text{Error máximo}$$

que son cantidades estrictamente positivas si algún $x_i \neq \bar{x}$.

2.2.1.2 Ejemplo

Obtener las desviaciones con respecto a la media en la siguiente distribución y comprobar que su suma es cero.

$l_{i-1} - l_i$	n_i
0 - 10	1
10 - 20	2
20 - 30	4
30 - 40	3

Solución:

$l_{i-1} - l_i$	n_i	x_i	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})n_i$
0 - 10	1	5	5	-19	-19
10 - 20	2	15	30	-9	-18
20 - 30	4	25	100	+1	+4
30 - 40	3	35	105	+11	+33
	$n=10$		$\sum x_i n_i = 240$		$\sum = 0$

La media aritmética es:

$$\bar{x} = \frac{1}{n} \sum x_i n_i = \frac{240}{10} = 24$$

Como se puede comprobar sumando los elementos de la última columna,

$$\sum (x_i - \bar{x}) \cdot n_i = 0$$

2.2.1.3 Proposición (König)

Para cualquier posible valor k que consideremos como candidato a medida central, \bar{x} lo mejora en el sentido de los mínimos cuadrados, es decir

$$\sum_{i=1}^n (x_i - \bar{x})^2 < \sum_{i=1}^n (x_i - k)^2 \quad \text{si } k \neq \bar{x}$$

Demostración

Sea $k \neq \bar{x}$. Veamos que el error cuadrático cometido por k es mayor que el de \bar{x} .

$$\begin{aligned} \sum_{i=1}^n (x_i - k)^2 &= \sum_{i=1}^n [x_i - (k - \bar{x} + \bar{x})]^2 && \text{(sumando y restando } \bar{x} \text{)} \\ &= \sum_{i=1}^n [(x_i - \bar{x}) - (k - \bar{x})]^2 && \text{(y usando el binomio de Newton...)} \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 - 2(\bar{x} - k) \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_0 + \underbrace{\sum_{i=1}^n (k - \bar{x})^2}_{n(k - \bar{x})^2 > 0} \\ &> \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

2.2.1.4 Proposición (Linealidad de la media)

$$Y = a + bX \implies \bar{y} = a + b\bar{x}$$

2.2.1.5 Proposición

Dados r grupos con n_1, n_2, \dots, n_r observaciones y siendo $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_r$ las respectivas medias de cada uno de ellos. Entonces la media de las $n = n_1 + \dots + n_r$ observaciones es

$$\bar{x} = \frac{n_1 \bar{x}_1 + \dots + n_r \bar{x}_r}{n_1 + \dots + n_r}$$

Demostración

Vamos a llamar x_{ij} a la j -ésima observación del grupo i ; Entonces tenemos

$$\left. \begin{array}{l} 1^{\text{er}} \text{ grupo} \rightarrow x_{11} \quad \dots \quad x_{1n_1} \\ 2^{\text{o}} \text{ grupo} \rightarrow x_{21} \quad \dots \quad x_{2n_2} \\ \dots \\ r^{\text{esimo}} \text{ grupo} \rightarrow x_{r1} \quad \dots \quad x_{rn_r} \end{array} \right\} \implies \begin{array}{l} \bar{x}_1 = \left(\sum_{j=1}^{n_1} x_{1j} \right) / n_1 \\ \bar{x}_2 = \left(\sum_{j=1}^{n_2} x_{2j} \right) / n_2 \\ \dots \\ \bar{x}_r = \left(\sum_{j=1}^{n_r} x_{rj} \right) / n_r \end{array}$$

Así, agrupando convenientemente las observaciones se llega a que

$$\begin{aligned}\bar{x} &= \frac{(x_{11} + \dots + x_{1n_1}) + (x_{22} + \dots + x_{2n_2}) + \dots + (x_{r1} + \dots + x_{rn_r})}{n_1 + n_2 + \dots + n_r} \\ &= \frac{n_1 \bar{x}_1 + \dots + n_r \bar{x}_r}{n}\end{aligned}$$

2.2.1.6 Observación

A pesar de las buenas propiedades que ofrece la media, ésta posee algunos inconvenientes:

Uno de ellos es que es muy sensible a los valores extremos de la variable: ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección. En consecuencia,

- no es recomendable usar la media como medida central en las distribuciones muy asimétricas;
- Depende de la división en intervalos en el caso de variables continuas.
- Si consideramos una variable discreta, por ejemplo, *el número de hijos en las familias de Málaga* el valor de la media puede no pertenecer al conjunto de valores de la variable; Por ejemplo $\bar{x} = 2'5$ hijos.

2.2.1.7 Cálculo abreviado

Se puede utilizar la [linealidad de la media](#) para simplificar las operaciones necesarias para su cálculo mediante un cambio de origen y de unidad de medida. El método consiste en lo siguiente:

1. Tomamos a un número que exprese aproximadamente el tipo de unidad con la que se trabaja. Por ejemplo, si las unidades que usamos son millones, tomamos $a=1.000.000$.
2. Seleccionamos un punto cualquiera de la zona central de la tabla, x_0 . Este punto jugará el papel de origen de referencia.
3. Cambiamos a la variable

$$\begin{aligned}Z &= \frac{X - x_0}{a} & \implies & \bar{z} = \frac{\bar{x} - x_0}{a} \\ & & \implies & \bar{x} = a\bar{z} + x_0\end{aligned}$$

4. Construimos de este modo la tabla de la variable Z, para la que es más fácil calcular \bar{z} directamente, y después se calcula \bar{x} mediante la relación (2.2).

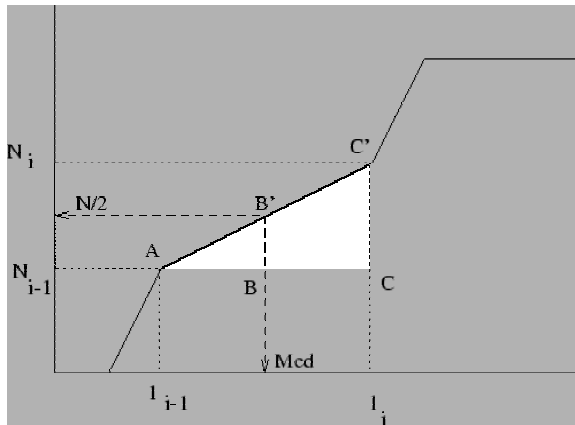
2.2.1.8 Medias generalizadas

En función del tipo de problema varias generalizaciones de la media pueden ser consideradas. He aquí algunas de ellas aplicadas a unas observaciones x_1, \dots, x_n :

2.2.2 La mediana

Consideramos una variable discreta X cuyas observaciones en una tabla estadística han sido ordenadas de menor a mayor. Llamaremos mediana, M_{ed} al primer valor de la variable que deja por debajo de sí al **50%** de las observaciones. Por tanto, si n es el número de observaciones, la mediana corresponderá a la observación $[n/2]+1$, donde representamos por $[\cdot]$ la parte entera de un número.

Figura: Cálculo geométrico de la mediana



En el caso de variables continuas, las clases vienen dadas por intervalos, y aquí la fórmula de la mediana se complica un poco más (pero no demasiado): Sea $(l_{i-1}, l_i]$ el intervalo donde hemos encontrado que por debajo están el **50%** de las observaciones. Entonces se obtiene la mediana a partir de las frecuencias absolutas acumuladas, mediante interpolación lineal (teorema de Thales) como sigue (figura 2.2):

$$\frac{CC'}{AC} = \frac{BB'}{AB} \implies \frac{n_i}{a_i} = \frac{\frac{n}{2} - N_{i-1}}{M_{ed} - l_{i-1}}$$

$$\implies \boxed{M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i}$$

2.2.2.1 Observación

La relación (2.2) corresponde a definir para cada posible observación, $x \in (l_{j-1}, l_j]$, su frecuencia relativa acumulada, $F(x)$, por interpolación lineal entre los valores $F(l_{j-1}) = F_{j-1}$ y $F(l_j) = F_j$ de forma que

$$F(x) = F(l_{j-1}) + \frac{F(l_j) - F(l_{j-1})}{a_j} (x - l_{j-1})$$

De este modo, M_{ed} es el punto donde $F(M_{ed}) = \frac{1}{2}$. Esto equivale a decir que la mediana divide al histograma en dos partes de áreas iguales a $\frac{1}{2}$.

2.2.2.2 Observación

Entre las propiedades de la mediana, vamos a destacar las siguientes:

- Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla.
- A diferencia de la media, la mediana de una variable discreta es siempre un valor de la variable que estudiamos (ej. La mediana de una variable *número de hijos* toma siempre valores enteros).

- Si una población está formada por 2 subpoblaciones de medianas M_{ed1} y M_{ed2} , sólo se puede afirmar que la mediana, M_{ed} , de la población está comprendida entre M_{ed1} y M_{ed2}

$$M_{ed1} \leq M_{ed} \leq M_{ed2}$$

- El mayor defecto de la mediana es que tiene unas propiedades matemáticas complicadas, lo que hace que sea muy difícil de utilizar en *inferencia estadística*.
- Es función de los intervalos escogidos.
- Puede ser calculada aunque el intervalo inferior o el superior no tenga límites.
- La suma de las diferencias de los valores absolutos de n puntuaciones respecto a su mediana es menor o igual que cualquier otro valor. Este es el equivalente al teorema de König (proposición 2.1) con respecto a la media, pero donde se considera como medida de dispersión a:

$$\sum_{i=1}^n |x_i - M_{ed}|$$

2.2.2.3 Ejemplo

Sea X una variable discreta que ha presentado sobre una muestra las modalidades

$$X \sim 2, 5, 7, 9, 12 \implies \bar{x} = 7, \quad M_{ed} = 7$$

Si cambiamos la última observación por otra anormalmente grande, esto no afecta a la mediana, pero sí a la media:

$$X \sim 2, 5, 7, 9, 125 \implies \bar{x} = 29,6; \quad M_{ed} = 7$$

En este caso la media no es un posible valor de la variable (discreta), y se ha visto muy afectada por la observación extrema. Este no ha sido el caso para la mediana.

2.2.2.4 Ejemplo

Obtener la media aritmética y la mediana en la distribución adjunta. Determinar gráficamente cuál de los dos promedios es más significativo.

$l_{i-1} - l_i$	n_i
0 - 10	60
10 - 20	80
20 - 30	30
30 - 100	20
100 - 500	10

Solución:

$l_{i-1} - l_i$	n_i	a_i	x_i	$x_i n_i$	N_i	n_i^f
0 - 10	60	10	5	300	60	60
10 - 20	80	10	15	1.200	140	80
20 - 30	30	10	25	750	170	30

30 - 100	20	70	65	1.300	190	2,9
100 - 500	10	400	300	3.000	200	0,25
	$n=200$			$\sum x_i n_i = 6.550$		

La media aritmética es:

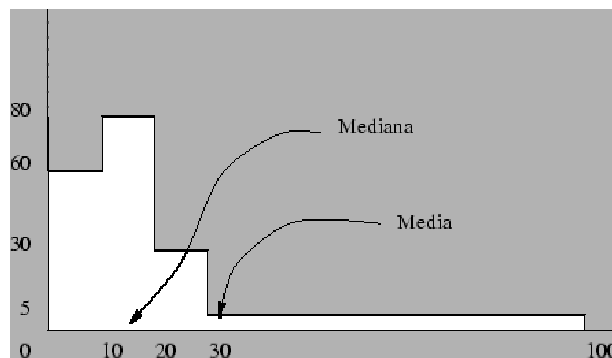
$$\bar{x} = \frac{1}{n} \sum x_i = \frac{6.550}{200} = 32,75$$

La primera frecuencia absoluta acumulada que supera el valor $n/2=100$ es $N_i=140$. Por ello el intervalo mediano es [10;20). Así:

$$M_{ed} = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \cdot a_i = 10 + \frac{100 - 60}{80} \times 10 = 15$$

Para ver la representatividad de ambos promedios, realizamos el histograma de la figura 2.3, y observamos que dada la forma de la distribución, la mediana es más representativa que la media.

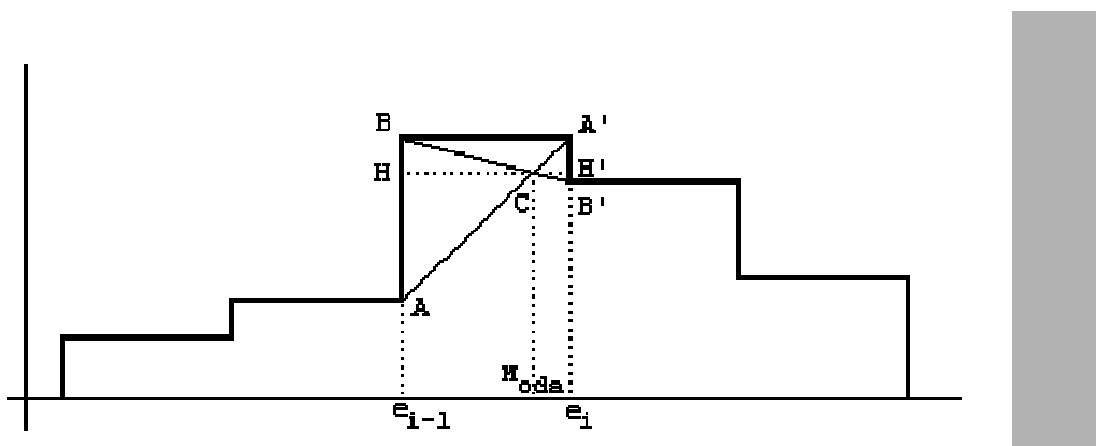
Figura: Para esta distribución de frecuencias es más útil usar como estadístico de tendencia central la mediana que la media.



2.2.3 La moda

Llamaremos moda a cualquier máximo relativo de la distribución de frecuencias, es decir, cualquier valor de la variable que posea una frecuencia mayor que su anterior y su posterior.

Figura: Cálculo geométrico de la moda



En el caso de variables continuas es más correcto hablar de intervalos modales. Una vez que este intervalo, $(l_{i-1}, l_i]$, se ha obtenido, se utiliza la siguiente fórmula para calcular la moda, que está motivada en la figura 2.4:

$$\frac{HC}{AB} = \frac{H'C}{A'B'} = \frac{HC + H'C}{AB + A'B'}$$

$$\Rightarrow \frac{M_{oda} - l_{i-1}}{n_i - n_{i-1}} = \frac{a_i}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

$$\Rightarrow \boxed{M_{oda} = l_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i}$$

Observación

De la moda destacamos las siguientes propiedades:

- Es muy fácil de calcular.
- Puede no ser única.
- Es función de los intervalos elegidos a través de su amplitud, número y límites de los mismos.
- Aunque el primero o el último de los intervalos no posean extremos inferior o superior respectivamente, la moda puede ser calculada.

2.2.4 Relación entre media, mediana y moda

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media).

En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

Veamos un ejemplo de cálculo de estas tres magnitudes.

Ejemplo

Consideramos una tabla estadística relativa a una variable continua, de la que nos dan los intervalos, las marcas de clase c_i , y las frecuencias absolutas, n_i .

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

Para calcular la media podemos añadir una columna con las cantidades $n_i c_i$. La suma de los términos de esa columna dividida por $n=12$ es la media:

Intervalos	c_i	n_i	N_i	$n_i c_i$
------------	-------	-------	-------	-----------

0 -- 2	1	2	2	2
2 -- 4	3	1	3	3
4 -- 6	5	4	7	20
6 -- 8	7	3	10	21
8 - 10	9	2	12	18
	12		64	

$$\bar{x} = \frac{64}{12} = 5,3$$

La mediana es el valor de la variable que deja por debajo de sí a la mitad de las n observaciones, es decir 6. Construimos la tabla de las frecuencias absolutas acumuladas, N_i , y vemos que eso ocurre en la modalidad tercera, es decir,

$$\begin{aligned}
 i &= 3 && \text{Observación} \\
 (l_{i-1}, l_i] &= (4; 6] && \text{Intervalo donde se encuentra la mediana} \\
 M_{ed} &= l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i = 4 + \frac{\frac{12}{2} - 3}{4} \cdot 2 = 5,5 \in (l_{i-1}, l_i]
 \end{aligned}$$

Para el cálculo de la **moda**, lo primero es encontrar los intervalos modales, buscando los máximos relativos en la columna de las frecuencias absolutas, n_i . Vemos que hay dos modas, correspondientes a las modalidades $i=1$, $i=3$. En el primer intervalo modal, $(l_{0,1})=(0,2]$, la moda se calcula como

$$M_{oda} = l_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i = 0 + \frac{2 - 0}{(2 - 0) + (2 - 1)} \cdot 2 = 1,3$$

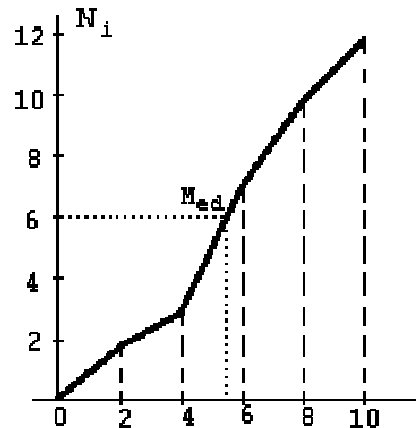
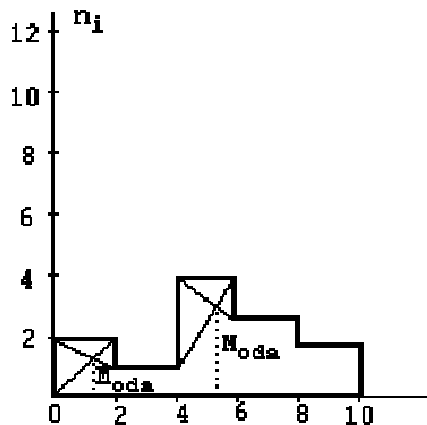
El segundo intervalo modal es $(l_{2,3})=(4;6]$, siendo la moda el punto perteneciente al mismo que se obtiene como:

$$M_{oda} = l_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i = 4 + \frac{4 - 1}{(4 - 1) + (4 - 3)} \cdot 2 = 5,5$$

En este caso, como se ve en la figura 2.5, la moda no toma un valor único, sino el conjunto

$$M_{oda} = \{1,3; 5,5\}$$

Figura: Diagramas diferencial e integral con cálculo geométrico de la moda y de la mediana de la variable.



2.3 Estadísticos de posición

Para una variable discreta, se define el percentil de orden k , como la observación, P_k , que deja por debajo de sí el $k\%$ de la población. Esta definición nos recuerda a la mediana, pues como consecuencia de la definición es evidente que

$$M_{ed} = P_{50}$$

En el caso de una variable continua, el intervalo donde se encuentra $P_k \in (l_{i-1}, l_i]$, se calcula buscando el que deja debajo de sí al $k\%$ de las observaciones. Dentro de él, P_k se obtiene según la relación:

$$P_k = l_{i-1} + \frac{n \frac{k}{100} - N_{i-1}}{n_i} \cdot a_i$$

Por su propia naturaleza, el percentil puede estar situado en cualquier lugar de la distribución, por lo que no puede considerarse como una medida de tendencia central.

Los cuartiles, Q_i , son un caso particular de los percentiles. Hay 3, y se definen como:

$$\begin{aligned} Q_1 &= P_{25} \\ Q_2 &= P_{50} = M_{ed} \\ Q_3 &= P_{75} \end{aligned}$$

De forma análoga se definen los deciles como los valores de la variable que dividen a las observaciones en 10 grupos de igual tamaño. Más precisamente, definimos D_1, D_2, \dots, D_9 como:

$$D_i = P_{10i} \quad i = 1, \dots, 9$$

Los percentiles (que incluyen a la mediana, cuartiles y deciles) también son denominados estadísticos de posición.

2.3.1 Ejemplo

Dada la siguiente distribución en el número de hijos de cien familias, calcular sus cuartiles.

x_i	n_i	N_i
-------	-------	-------

0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
	n=100	

Solución:

1. Primer cuartil:

$$\frac{n}{4} = 25; \text{ Primera } N_i > n/4 = 39; \text{ luego } Q_1 = 2.$$

2. Segundo cuartil:

$$\frac{2n}{4} = 50; \text{ Primera } N_i > 2n/4 = 65; \text{ luego } Q_2 = 3.$$

3. Tercer cuartil:

$$\frac{3n}{4} = 75; \text{ Primera } N_i > 3n/4 = 85; \text{ luego } Q_3 = 4.$$

2.3.2 Ejemplo

Calcular los cuartiles en la siguiente distribución de una variable continua:

$l_{i-1} - l_i$	n_i	N_i
0 - 1	10	10
1 - 2	12	22
2 - 3	12	34
3 - 4	10	44
4 - 5	7	51
	n=51	

Solución:

1. Primer cuartil

$$\frac{N}{4} = 12,75; \text{ Primera } N_i > n/4 = 22; \text{ La línea } i \text{ es la del intervalo } [1; 2)$$

$$Q_1 = l_{i-1} + \frac{\frac{n}{4} - N_{i-1}}{n_i} a_i = 1 + \frac{12,75 - 10}{12} \times 1 = 1,23$$

2. Segundo cuartil:

$$\frac{2n}{4} = 25,5; \text{ Primera } N_i > 2n/4 = 34; \text{ La línea } i \text{ es la del intervalo } [2;3)$$

$$Q_2 = l_{i-1} + \frac{\frac{2n}{4} - N_{i-1}}{n_i} a_i = 2 + \frac{25,5 - 22}{12} \times 1 = 2,29$$

3. Tercer cuartil

$$\frac{3n}{4} = 38,25; \text{ Primera } N_i > 3n/4 = 44; \text{ La línea } i \text{ es la del intervalo } [3;4)$$

$$Q_3 = l_{i-1} + \frac{\frac{3n}{4} - N_{i-1}}{n_i} a_i = 3 + \frac{38,25 - 34}{10} \times 1 = 3,445$$

2.3.3 Ejemplo

Han sido ordenados los pesos de 21 personas en la siguiente tabla:

Intervalos	f.a.
$l_{i-1} - l_i$	n_i
38 -- 45	3
45 -- 52	2
52 -- 59	7
59 -- 66	3
66 -- 73	6
	21

Encontrar aquellos valores que dividen a los datos en 4 partes con el mismo número de observaciones.

Solución: Las cantidades que buscamos son los tres cuartiles: Q_1 , Q_2 y Q_3 . Para calcularlos, le añadimos a la tabla las columnas con las frecuencias acumuladas, para localizar qué intervalos son los que contienen a los cuartiles buscados:

$l_{i-1} - l_i$	n_i	N_i	
38 -- 45	3	3	
45 -- 52	2	5	
52 -- 59	7	12	$\ni Q_1, Q_2$
59 -- 66	3	15	
66 -- 73	6	21	$\ni Q_3$
	21		

Q_1 Q_2
v se encuentran en

el intervalo 52--59, ya que $N_3=12$ es la primera f.a.a.

que supera a $21 \cdot 1/4$ y $21 \cdot 2/4$

Q_3 está en 66--73, pues $N_5=21$ es el primer N_i mayor que $21 \cdot 3/4$

Así se tiene que:

$$\begin{aligned} \frac{1}{4} \cdot 21 = 5,25 \Rightarrow i = 3 \Rightarrow Q_1 &= l_{i-1} + \frac{\frac{1}{4}n - N_{i-1}}{n_i} \cdot a_i \\ &= 52 + \frac{5,25 - 5}{7} \cdot 7 = 52,25 \end{aligned}$$

$$\begin{aligned} \frac{2}{4} \cdot 21 = 10,5 \Rightarrow i = 3 \Rightarrow Q_2 &= l_{i-1} + \frac{\frac{2}{4}n - N_{i-1}}{n_i} \cdot a_i \\ &= 52 + \frac{10,5 - 5}{7} \cdot 7 = 57,5 \end{aligned}$$

$$\begin{aligned} \frac{3}{4} \cdot 21 = 15,75 \Rightarrow i = 5 \Rightarrow Q_3 &= l_{i-1} + \frac{\frac{3}{4}n - N_{i-1}}{n_i} \cdot a_i \\ &= 66 + \frac{15,75 - 15}{6} \cdot 7 = 66,875 \end{aligned}$$

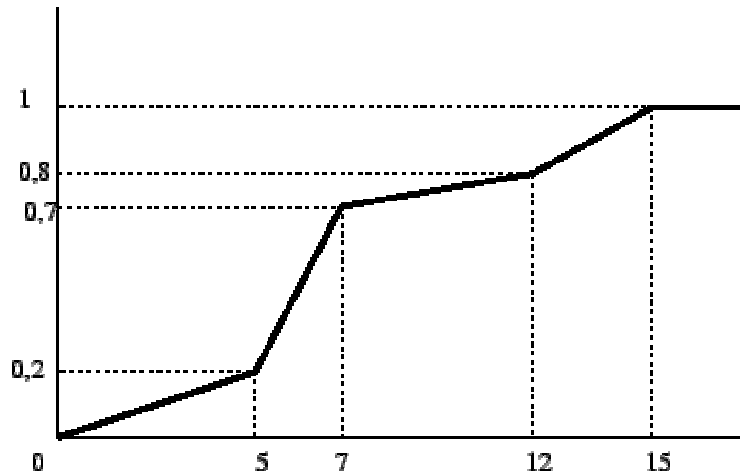
Obsérvese que $Q_2 = M_{med}$. Esto es lógico, ya que la mediana divide a la distribución en dos partes con el mismo número de observaciones, y Q_2 , hace lo mismo, pues es deja a dos cuartos de los datos por arriba y otros dos cuartos por abajo.

2.3.4 Ejemplo

La distribución de una variable tiene por polígono acumulado de frecuencias el de la figura 2.6. Si el número total de observaciones es 50:

1. Elaborar una tabla estadística con los siguientes elementos: intervalos, marcas de clase, frecuencia absoluta, frecuencia absoluta acumulada, frecuencias relativa y frecuencias relativa acumulada.
2. Cuántas observaciones tuvieron un valor inferior a 10, cuántas inferior a 8 y cuántas fueron superior a 11.
3. Calcule las modas.
4. Determine los cuartiles.

Figura: Diagrama acumulado de frecuencias relativas.



Solución:

1. En la siguiente tabla se proporciona la información pedida y algunos cálculos auxiliares que nos permitirán responder a otras cuestiones.

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	$n_i \cdot f'$
0 - 5	10	10	0,2	0,3	2,5	5	2
5 - 7	25	35	0,5	0,7	6	2	12,5
7 - 12	5	40	0,1	0,8	9,5	5	1
12 - 15	10	50	0,2	1	13,5	7	3,33

2. Calculemos el número de observaciones pedido:

$$\begin{array}{l} 7 \text{ a } 12 \\ 7 \text{ a } 10 \end{array} \begin{array}{l} \text{-----} 5 \\ \text{-----} x \end{array} \Leftrightarrow \begin{array}{l} 5 \text{-----} 5 \\ 3 \text{-----} x \end{array} \Rightarrow x = \frac{3 \times 5}{5} = 3$$

10 + 25+3 = 38 observaciones tomaron un valor inferior a 10

$$\begin{array}{l} 7 \text{ a } 12 \\ 7 \text{ a } 8 \end{array} \begin{array}{l} \text{-----} 5 \\ \text{-----} x \end{array} \Leftrightarrow \begin{array}{l} 5 \text{-----} 5 \\ 1 \text{-----} x \end{array} \Rightarrow x = \frac{1 \times 5}{5} = 1$$

10 + 25+1 = 36 observaciones tomaron un valor inferior a 8

$$\begin{array}{l} 7 \text{ a } 12 \\ 7 \text{ a } 11 \end{array} \begin{array}{l} \text{-----} 5 \\ \text{-----} x \end{array} \Leftrightarrow \begin{array}{l} 5 \text{-----} 5 \\ 4 \text{-----} x \end{array} \Rightarrow x = \frac{4 \times 5}{5} = 4$$

50 -(10 + 25+4) = 50-39=11 observaciones tomaron un valor superior a 11

3. Hay dos modas. Calculemos la más representativa:

$$M_{oda} = l_{i-1} + \frac{n_{i+1}'}{n_{i-1}' + n_{i+1}'} \cdot a_i = 5 + \frac{1}{2+1} \cdot 2 = 5,66$$

4. Cuartiles:

$$Q_1 = l_{i-1} + \frac{n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{12,5 - 10}{25} \cdot 2 = 5,2$$

$$Q_2 = l_{i-1} + \frac{2n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{25 - 10}{25} \cdot 2 = 6,2$$

$$Q_3 = l_{i-1} + \frac{3n/4 - N_{i-1}}{n_i} \cdot a_i = 7 + \frac{37,5 - 35}{5} \cdot 5 = 9,5$$

2.4 Medidas de variabilidad o dispersión

Los estadísticos de tendencia central o posición nos indican donde se sitúa un grupo de puntuaciones. Los de variabilidad o dispersión nos indican si esas puntuaciones o valores están próximas entre sí o si por el contrario están o muy dispersas.

Una medida razonable de la variabilidad podría ser la amplitud o rango, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto. Es fácil de calcular y sus unidades son las mismas que las de la variable, aunque posee varios inconvenientes:

- No utiliza todas las observaciones (sólo dos de ellas);
- Se puede ver muy afectada por alguna observación extrema;
- El rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

En el transcurso de esta sección, veremos medidas de dispersión mejores que la anterior. Estas se determinan en función de la distancia entre las observaciones y algún estadístico de tendencia central.

2.4.1 Desviación media, D_m

Se define la desviación media como la media de las diferencias en valor absoluto de los valores de la variable a la media, es decir, si tenemos un conjunto de n observaciones, x_1, \dots, x_n , entonces

$$D_m = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Si los datos están agrupados en una tabla estadística es más sencillo usar la relación

$$D_m = \frac{1}{n} \sum_{i=1}^k |x_i - \bar{x}| n_i$$

Como se observa, la desviación media guarda las mismas dimensiones que las observaciones. La suma de valores absolutos es relativamente sencilla de calcular, pero esta simplicidad tiene un inconveniente: Desde el punto de vista geométrico, la distancia que induce la desviación media en el espacio de observaciones no es la natural (no permite definir ángulos entre dos conjuntos de observaciones). Esto hace que sea muy engorroso trabajar con ella a la hora de hacer inferencia a la población.

2.4.2 Varianza y desviación típica

Como forma de medir la dispersión de los datos hemos descartado:

- $\sum_{i=1}^n (x_i - \bar{x})$, pues sabemos que esa suma vale 0, ya que las desviaciones con respecto a la media se compensan al haber términos en esa suma que son de signos distintos.
- Para tener el mismo signo al sumar las desviaciones con respecto a la media podemos realizar la suma con valores absolutos. Esto nos lleva a la D_m , pero como hemos mencionado, tiene poco interés por las dificultades que presenta.

Si las desviaciones con respecto a la media las consideramos al cuadrado, $(x_i - \bar{x})^2$, de nuevo obtenemos que todos los sumandos tienen el mismo signo (positivo). Esta es además la forma de medir la dispersión de los datos de forma que sus propiedades matemáticas son más fáciles de utilizar. Vamos a definir entonces dos estadísticos que serán fundamentales en el resto del curso: La varianza y la desviación típica.

La varianza, S^2 , se define como la media de las diferencias cuadráticas de n puntuaciones con respecto

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

a su media aritmética, es decir

Para datos agrupados en tablas, usando las notaciones establecidas en los capítulos anteriores, la

$$S^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^2 n_i$$

varianza se puede escribir como

Una fórmula equivalente para el cálculo de la varianza está basada en lo siguiente:

$$\begin{aligned} S^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \underbrace{\frac{1}{n} \sum_{i=1}^n (x_i)}_{=\bar{x}} + \frac{1}{n} n \bar{x}^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \end{aligned}$$

Con lo cual se tiene

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Si los datos están agrupados en tablas, es evidente que

$$S^2 = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - \bar{x}^2$$

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en **metros²**). Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la desviación típica, S , como

$$S = \sqrt{S^2}$$

2.4.2.1 Ejemplo

Calcular la varianza y desviación típica de las siguientes cantidades medidas en metros:

3,3,4,4,5

Solución: Para calcular dichas medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a medir las diferencias. Éste es la media:

$$\bar{x} = (3 + 3 + 4 + 4 + 5)/5 = 3,8 \text{ metros}$$

La varianza es:

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{5} (3^2 + 3^2 + 4^2 + 4^2 + 5^2) - 3,8^2 = 0,56 \text{ metros}^2$$

siendo la desviación típica su raíz cuadrada:

$$S = \sqrt{S^2} = \sqrt{0,56} = 0,748 \text{ metros}$$

Las siguientes propiedades de la varianza (respectivamente, desviación típica) son importantes a la hora de hacer un cambio de origen y escala a una variable. En primer lugar, la varianza (resp. Desviación típica) no se ve afectada si al conjunto de valores de la variable se le añade una constante. Si además cada observación es multiplicada por otra constante, en este caso la varianza cambia en relación al cuadrado de la constante (resp. La desviación típica cambia en relación al valor absoluto de la constante). Esto queda precisado en la siguiente proposición:

2.4.2.2 Proposición

Si $Y = aX + b$ entonces $S^2_Y = a^2 S^2_X$

Demostración

Para cada observación x_i de X , $i = 1, \dots, n$, tenemos una observación de Y que es por definición $y_i = ax_i + b$. Por la proposición [2.1](#), se tiene que $\bar{y} = a\bar{x} + b$. Por tanto, la varianza de Y es

$$\begin{aligned}
 S^2_Y &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\
 &= \frac{1}{n} \sum_{i=1}^n \left[\underbrace{(ax_i + b)}_{y_i} - \underbrace{(a\bar{x} + b)}_{\bar{y}} \right]^2 \\
 &= \frac{1}{n} \sum_{i=1}^n a^2 (x_i - \bar{x})^2 \\
 &= a^2 S^2_X
 \end{aligned}$$

2.4.2.3 Observación

Las consecuencias del anterior resultado eran de esperar: Si los resultados de una medida son trasladados una cantidad b , la dispersión de los mismos no aumenta. Si estos mismos datos se multiplican por una cantidad $a < 1$, el resultado tenderá a concentrarse alrededor de su media (menor varianza). Si por el contrario $a > 1$ habrá mayor dispersión.

Otra propiedad fundamental de la varianza es la siguiente:

2.4.2.4 Proposición

Dados r grupos, cada uno de ellos formado por n_i observaciones de media \bar{x}_i y de varianza S_i^2 . Entonces la varianza, S^2 , del conjunto de todas las $n = n_1 + \dots + n_r$ observaciones vale

$$S^2 = \frac{1}{n} \sum_{i=1}^r n_i S_i^2 + \frac{1}{n} \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2$$

Demostración

Dicho de otro modo, pretendemos demostrar que la varianza total es igual a la media de las varianzas más la varianza de las medias. Comenzamos denotando mediante x_{ij} la observación j -ésima en el i -ésimo grupo, donde $i = 1, \dots, r$ y $j = 1, \dots, n_i$. Entonces

$$\begin{aligned}
 S^2 &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \\
 &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})]^2 \\
 &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + 2 \sum_{i=1}^r (\bar{x}_i - \bar{x}) \underbrace{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)}_{=0} + \frac{1}{n} \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2 \\
 &= \frac{1}{n} \sum_{i=1}^r n_i S_i^2 + 0 + \frac{1}{n} \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2
 \end{aligned}$$

2.4.2.5 Observación

Además de las propiedades que hemos demostrado sobre la varianza (y por tanto sobre la desviación típica), será conveniente tener siempre en mente otras que enunciaremos a continuación:

Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza. La razón es que si miramos su definición, la varianza es función de cada una de las puntuaciones.

- Si se calculan a través de los datos agrupados en una tabla, dependen de los intervalos elegidos. Es decir, cometemos cierto error en el cálculo de la varianza cuando los datos han sido resumidos en una tabla estadística mediante intervalos, en lugar de haber sido calculados directamente como datos no agrupados. Este error no será importante si la elección del número de intervalos, amplitud y límites de los mismos ha sido adecuada.
- La desviación típica tiene la propiedad de que en el intervalo

$$(\bar{x} - 2S, \bar{x} + 2S) \stackrel{\text{def}}{\sim} \bar{x} \pm 2S$$

se encuentra, al menos, el 75% de las observaciones (vease más adelante el teorema de Thebycheff). Incluso si tenemos muchos datos y estos provienen de una distribución normal (se definirá este concepto más adelante), podremos llegar al **95%**.

No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central.

2.4.2.6 Método abreviado para el cálculo de la varianza

La proposición de la página puede ser utilizada para simplificar cálculos al igual que vimos en el ejemplo [2.1](#). Si una variable X toma unos valores para los cuales las operaciones de cálculo de media y varianza son tediosas, podemos realizar los cálculos sobre una variable Z definida como

$$Z = \frac{X - x_0}{a}$$

Una vez que han sido calculadas \bar{z} y S_Z^2 , obtenemos \bar{x} y S_X^2 teniendo en cuenta que:

$$X = aZ + x_0 \implies \begin{cases} \bar{x} = a\bar{z} + x_0 \\ S_X^2 = a^2 S_Z^2 \end{cases}$$

2.4.2.7 Grados de libertad

Los grados de libertad de un estadístico calculado sobre n datos se refieren al número de cantidades independientes que se necesitan en su cálculo, menos el número de restricciones que ligan a las observaciones y el estadístico. Es decir, normalmente n-1.

Ejemplo. Consideramos una serie de valores de una variable,

$$x_i \rightsquigarrow 2, 5, 7, 9, 12$$

que han sido tomados de forma independiente.

Su media es $\bar{x} = 7$ y se ha calculado a partir de las $n=5$ observaciones independientes x_i , que están ligadas a la media por la relación:

$$\bar{x} = \frac{1}{n} \sum x_i$$

Luego el número de grados de libertad de la media es $n-1=4$.

Si calculamos a continuación la varianza, se han de sumar n cantidades

$$\frac{(x_i - \bar{x})^2}{n}$$

Sin embargo esas cantidades no son totalmente independientes, pues están ligadas por una restricción:

$$\sum_{i=1}^n (x_i - (\sum_{i=1}^n x_i) / n) = 0$$

El número de grados de libertad del estadístico es el número de observaciones de la variable menos el número de restricciones que verifican, así que en este caso, los grados de libertad de la varianza sobre los $n=5$ datos son también $n-1 = 4$.

Un principio general de la teoría matemática nos dice que si pretendemos calcular de modo aproximado la varianza de una población a partir de la varianza de una muestra suya, se tiene que el error cometido es generalmente más pequeño, si en vez de considerar como estimación de la varianza de la población, a la varianza muestral

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

consideramos lo que se denomina cuasivarianza muestral, \hat{s}^2 que se calcula como la anterior, pero cambiando el denominador por el número de grados de libertad, $n-1$:

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n s^2}{n-1}$$

Sobre este punto incidiremos más adelante, ya que es fundamental en estadística inferencial.

2.4.2.8 Tipificación

Se conoce por tipificación al proceso de restar la media y dividir por su desviación típica a una variable X . De este modo se obtiene una nueva variable

$$Z = \frac{X - \bar{x}}{s}$$

de media $\bar{z} = 0$ y desviación típica $s_z = 1$, que denominamos variable tipificada.

Esta nueva variable carece de unidades y permite hacer comparables dos medidas que en un principio no lo son, por aludir a conceptos diferentes. Así por ejemplo nos podemos preguntar si un elefante es más grueso que una hormiga determinada, cada uno en relación a su población. También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes. Por ejemplo si deseamos comparar el nivel académico de dos estudiantes de diferentes Universidades para la concesión de una beca de estudios, en principio sería injusto concederla directamente al que posea una

nota media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro. En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad.

2.4.3 Coeficiente de variación

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones. Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de elefantes de dos circos diferentes, S nos dará información útil.

¿Pero qué ocurre si lo que comparamos es la altura de unos elefantes con respecto a su peso? Tanto la media como la desviación típica, \bar{x} y S , se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo de que una de las medidas sea de longitud y la otra sea de masa. El mismo problema se plantea si medimos cierta cantidad, por ejemplo la masa, de dos poblaciones, pero con distintas unidades. Este es el caso en que comparamos el peso en toneladas de una población de 100 elefantes con el correspondiente en miligramos de una población de 50 hormigas.

El problema no se resuelve tomando las mismas escalas para ambas poblaciones. Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que los elefantes (toneladas). Si la ingeniería genética no nos sorprende con alguna barbaridad, lo lógico es que la dispersión de la variable peso de las hormigas sea prácticamente nula (¡Aunque haya algunas que sean 1.000 veces mayores que otras!)

En los dos primeros casos mencionados anteriormente, el problema viene de la dimensionalidad de las variables, y en el tercero de la diferencia enorme entre las medias de ambas poblaciones. El coeficiente de variación es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

$$CV = \frac{S_x}{\bar{x}}$$

Basta dar una rápida mirada a la definición del coeficiente de variación, para ver que las siguientes consideraciones deben ser tenidas en cuenta:

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que $\bar{x} > 0$.
- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, entonces $CV_Y < CV_X$, ya que la desviación típica no es sensible ante cambios de origen, pero si la media. Lo contrario ocurre si restamos ($b < 0$).

$$CV_Y = \frac{S_Y}{\bar{y}} = \frac{S_X}{\bar{x} + b} < \frac{S_X}{\bar{x}} = CV_X$$

- Es invariante a cambios de escala. Si multiplicamos X por una constante a , para obtener $Y = aX$, entonces

$$CV_Y = \frac{S_Y}{\bar{y}} = \frac{S_{aX}}{a\bar{x}} = \frac{aS_X}{a\bar{x}} = CV_X$$

2.4.3.1 Observación

Es importante destacar que los *coeficientes de variación* sirven para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones), mientras que si deseamos comparar a dos *individuos* de cada uno de esos conjuntos, es necesario usar los *valores tipificados*.

2.4.3.2 Ejemplo

Dada la distribución de edades (medidas en años) en un colectivo de 100 personas, obtener:

1. La variable tipificada Z .
2. Valores de la media y varianza de Z .
3. Coeficiente de variación de Z .

Horas trabajadas	Num. empleados
0 -- 4	47
4 -- 10	32
10 -- 20	17
20 -- 40	4
	100

Solución:

Para calcular la variable tipificada

$$Z = \frac{X - \bar{x}}{S_X},$$

partimos de los datos del enunciado. Será necesario calcular en primer lugar la media y desviación típica de la variable original (X = años).

$l_{i-1} - l_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
0 -- 4	2	47	94	188
4 -- 10	7	32	224	1.568
10 -- 20	15	17	255	3.825
20 -- 40	30	4	120	3.600
		$n=100$	693	9.181

$$\bar{x} = \frac{693}{100} = 6,93 \text{ años}$$

$$S_X^2 = \frac{9.181}{100} - 6,93^2 = 43,78 \text{ años al cuadrado}$$

$$S_X = \sqrt{43,78} = 6,6 \text{ años}$$

A partir de estos valores podremos calcular los valores tipificados para las marcas de clase de cada intervalo y construir su distribución de frecuencias:

$$z_1 = \frac{2 - 6,93}{6,6} = -0,745$$

$$z_2 = \frac{7 - 6,93}{6,6} = 0,011$$

$$z_3 = \frac{15 - 6,93}{6,6} = 1,22$$

$$z_4 = \frac{30 - 6,93}{6,6} = 3,486$$

z_i	n_i	$z_i n_i$	$z_i^2 n_i$
-0,745	47	-35,015	26,086
0,011	32	0,352	0,004
1,220	17	20,720	25,303
3,486	4	13,944	48,609
	$n=100$	0,021	100,002

$$\bar{z} = \frac{0,021}{100} \approx 0$$

$$s_z^2 = \frac{100,02}{100} - 0^2 \approx 1$$

$$s_x = \sqrt{1} = 1$$

A pesar de que no se debe calcular el coeficiente de variación sobre variables que presenten valores negativos (y Z los presenta), lo calculamos con objeto de ilustrar el porqué:

$$CV = \frac{s_z}{\bar{z}} = \frac{1}{0} = \infty$$

Es decir, el coeficiente de variación no debe usarse nunca con variables tipificadas.

2.5 Problemas

Ejercicio 2.1. En el siguiente conjunto de números, se proporcionan los pesos (redondeados a la libra más próxima) de los bebés nacidos durante un cierto intervalo de tiempo en un hospital:

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 frecuencias 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 de la parte a.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?
7. Calcular las medidas de tendencia central.
8. Calcular las medidas de dispersión.
9. Calcular las medidas de forma.
10. ¿Es esta una distribución sesgada? De ser así, ¿en qué dirección?

11. Encontrar el percentil 24.

Ejercicio 2.2. A continuación se dan los resultados obtenidos con una muestra de 50 universitarios. la característica es el tiempo de reacción ante un estímulo auditivo:

0,110	0,110	0,126	0,112	0,117	0,113	0,135	0,107	0,122
0,113	0,098	0,122	0,105	0,103	0,119	0,100	0,117	0,113
0,124	0,118	0,132	0,108	0,115	0,120	0,107	0,123	0,109
0,117	0,111	0,112	0,101	0,112	0,111	0,119	0,103	0,100
0,108	0,120	0,099	0,102	0,129	0,115	0,121	0,130	0,134
0,118	0,106	0,128	0,094	0,1114				

1. ¿Cuál es la amplitud total de la distribución de los datos?
2. Obtenga la distribución de frecuencias absolutas y relativas.
3. Obtenga la distribución de frecuencias acumuladas, absolutas y relativas, con los intervalos anteriores.
4. Calcular la media y la varianza con los intervalos del apartado **b** y después calculense las mismas magnitudes sin ordenar los datos en una tabla estadística. ¿Con qué método se obtiene mayor precisión?
5. Dibuje el polígono de frecuencias relativas.
6. Dibuje el polígono de frecuencias relativas acumuladas.

Ejercicio 2.3. Con el fin de observar la relación entre la inteligencia y el nivel socioeconómico (medido por el salario mensual familiar) se tomaron dos grupos, uno formado con sujetos de cociente intelectual inferior a 95 y otro formado por los demás; De cada sujeto se anotó el salario mensual familiar. Teniendo en cuenta los resultados que se indican en la tabla:

Nivel socioeconómico	Sujetos con $CI < 95$	Sujetos con $CI \geq 95$
Intervalos	Frecuencia	Frecuencia
10 o menos $\equiv (4,10]$	75	19
10 - 16	35	26
16 - 22	20	25
22 - 28	30	30
28 - 34	25	54
más de 34 $\equiv (34,40]$	15	46

1. Dibuje un gráfico que permita comparar ambos grupos.
2. Calcule las medidas de tendencia central para aquellos sujetos con $CI < 95$.
3. Calcular las medidas de dispersión para aquellos sujetos con $CI \geq 95$.

Ejercicio 2.4. Un estudio consistió en anotar el número de palabras leídas en 15 segundos por un grupo de 120 sujetos disléxicos y 120 individuos normales. Teniendo en cuenta los resultados de la tabla

Nº de palabras leídas	Disléxicos n_D	Normales n_N
25 o menos $\equiv 25$	56	1
26	24	9
27	16	21
28	12	29

29	10	28
30 o más $\equiv 30$	2	32

calcule:

1. Las medias aritméticas de ambos grupos.
2. Las medianas de ambos grupos.
3. El porcentaje de sujetos disléxicos que superaron la mediana de los normales.
4. Compare la variabilidad relativa de ambos grupos.

Ejercicio 2.5. La tabla siguiente muestra la composición por edad, sexo y trabajo de un grupo de personas con tuberculosis pulmonar en la provincia de Vizcaya en el año 1979:

Edad	Trabajadores			No trabajadores			Totales		
	Varón	Mujer	Total	Varón	Mujer	Total	Varón	Mujer	Total
14-19	2	1	3	25	40	65	27	41	68
19-24	10	4	14	20	36	56	30	40	70
24-29	32	10	42	15	50	65	47	60	107
29-34	47	12	59	13	34	47	60	46	106
34-39	38	8	46	10	25	35	48	33	81
39-44	22	4	26	7	18	25	29	22	51

1. Representar gráficamente la distribución de frecuencias de aquellas personas trabajadoras que padecen tuberculosis.
2. Representar gráficamente la distribución de frecuencias de los varones no trabajadores que padecen tuberculosis.
3. Representar gráficamente la distribución de frecuencias del número total de mujeres que padecen tuberculosis.
4. ¿Cuál es la edad en la que se observa con mayor frecuencia que no trabajan los varones? ¿Y las mujeres? Determinar asimismo la edad más frecuente (sin distinción de sexos ni ocupación).
5. ¿Por debajo de qué edad está el 50% de los varones?
6. ¿Por encima de qué edad se encuentra el 80% de las mujeres?
7. Obtener la media, mediana y desviación típica de la distribución de las edades de la muestra total.
8. Estudiar la asimetría de las tres distribuciones.

Ejercicio 2.6. En una epidemia de escarlatina, se ha recogido el número de muertos en 40 ciudades de un país, obteniéndose la siguiente tabla:

Nº de muertos	0	1	2	3	4	5	6	7
Ciudades	7	11	10	7	1	2	1	1

1. Representar gráficamente estos datos.
2. Obtener la distribución acumulada y representarla.
3. Calcular media, mediana y moda.
4. Calcular la varianza y la desviación típica.
5. Porcentaje de ciudades con al menos 2 muertos.
6. Porcentaje de ciudades con más de 3 muertos.
7. Porcentaje de ciudades con a lo sumo 5 muertos.

3. Variables bidimensionales

3.1 Introducción

En lo estudiado anteriormente hemos podido aprender cómo a partir de la gran cantidad de datos que describen una muestra mediante una variable, X , se representan gráficamente los mismos de modo que resulta más intuitivo hacerse una idea de como se distribuyen las observaciones.

Otros conceptos que según hemos visto, también nos ayudan en el análisis, son los estadísticos de tendencia central, que nos indican hacia donde tienden a agruparse los datos (en el caso en que lo hagan), y los estadísticos de dispersión, que nos indican si las diferentes modalidades que presenta la variable están muy agrupadas alrededor de cierto valor central, o si por el contrario las variaciones que presentan las modalidades con respecto al valor central son grandes.

También sabemos determinar ya si los datos se distribuyen de forma simétrica a un lado y a otro de un valor central.

En este capítulo pretendemos estudiar una situación muy usual y por tanto de gran interés en la práctica:

Si Y es otra variable definida sobre la misma población que X , ¿será posible determinar si existe alguna relación entre las modalidades de X y de Y ?

Un ejemplo trivial consiste en considerar una población formada por alumnos de primero de Medicina y definir sobre ella las variables

X \equiv altura medida en centímetros,

Y \equiv altura medida en metros,

ya que la relación es determinista y clara: $Y=X/100$. Obsérvese que aunque la variable Y , como tal puede tener cierta dispersión, vista como función de X , su dispersión es nula.

Un ejemplo más parecido a lo que nos interesa realmente lo tenemos cuando sobre la misma población definimos las variables

X \equiv altura medida en centímetros,

Y \equiv peso medida en kilogramos.

Intuitivamente esperamos que exista cierta relación entre ambas variables, por ejemplo,

$Y = X - 110 \pm$ dispersión que nos expresa que (en media) a mayor altura se espera mayor peso. La relación no es exacta y por ello será necesario introducir algún término que exprese la dispersión de Y con respecto a la variable X .

Es fundamental de cara a realizar un trabajo de investigación experimental, conocer muy bien las técnicas de estudio de variables bidimensionales (y n -dimensionales en general). Baste para ello pensar que normalmente las relaciones entre las variables no son tan evidentes como se mencionó arriba. Por ejemplo:

¿Se puede decir que en un grupo de personas existe alguna relación entre X = tensión arterial e Y = edad?

Aunque en un principio la notación pueda resultar a veces algo desagradable, el lector podrá comprobar, al final del capítulo, que es bastante accesible. Por ello le pedimos que no se asuste. Al final verá que no son para tanto.

3.2 Tablas de doble entrada

Consideramos una población de n individuos, donde cada uno de ellos presenta dos caracteres que representamos mediante las variables X e Y . Representamos mediante

$$X \sim x_1, x_2, \dots, x_i, \dots, x_k$$

las k modalidades que presenta la variable X , y mediante

$$Y \sim y_1, y_2, \dots, y_j, \dots, y_p$$

las p modalidades de Y .

Con la intención de reunir en una sólo estructura toda la información disponible, creamos una tabla formada por $k \cdot p$ casillas, organizadas de forma que se tengan k filas y p columnas. La casilla denotada de forma general mediante el **subíndice** n_{ij} hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades x_i e y_j .

Y	y_1	y_2	...	y_j	...	y_p	
X							
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}	$n_{2\cdot}$
...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ip}	$n_{i\cdot}$
...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kp}	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$...	$n_{\cdot j}$...	$n_{\cdot p}$	$n_{\cdot \cdot}$

De este modo, para $i = 1, \dots, k$, $j = 1, \dots, p$, se tiene que n_{ij} es el número de individuos o frecuencia absoluta, que presentan a la vez las modalidades x_i e y_j .

El número de individuos que presentan la modalidad x_i , es lo que llamamos frecuencia absoluta marginal de x_i y se representa como **$n_{i\cdot}$** . Es evidente la igualdad

$$n_{i\cdot} = n_{i1} + n_{i2} + \dots + n_{ip} = \sum_{j=1}^p n_{ij}$$

Obsérvese que hemos escrito un símbolo **\cdot** en la "parte de las jotas" que simboliza que estamos considerando los elementos que presentan la modalidad x_i , independientemente de las modalidades que presente la variable Y . De forma análoga se define la frecuencia absoluta marginal de la modalidad y_j como

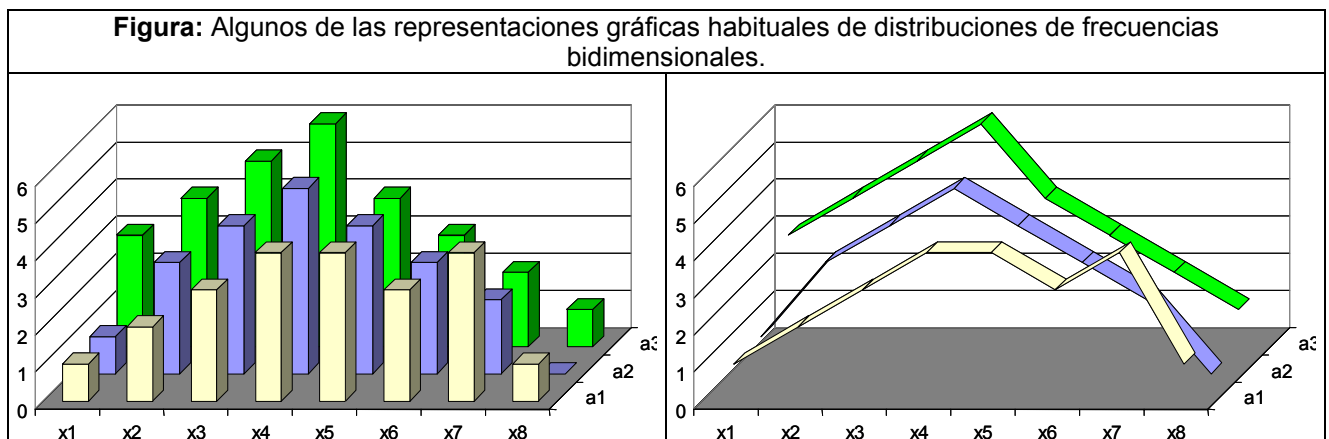
$$n_{\cdot j} = n_{1j} + n_{2j} + \dots + n_{kj} = \sum_{i=1}^k n_{ij}$$

Estas dos distribuciones de frecuencias $n_{i\cdot}$ para $i = 1, \dots, k$, y $n_{\cdot j}$ para $j = 1, \dots, p$ reciben el nombre de distribuciones marginales de X e Y respectivamente.

El número total de elementos de la población (o de la muestra), n lo obtenemos de cualquiera de las siguientes formas, que son equivalentes:

$$n = n_{\cdot\cdot} = \sum_{i=1}^k n_{i\cdot} = \sum_{j=1}^p n_{\cdot j} = \sum_{i=1}^k \sum_{j=1}^p n_{ij}$$

Las distribuciones de frecuencias de las variables bidimensionales también pueden ser representadas gráficamente. Al igual que en el caso unidimensional existen diferentes tipos de representaciones gráficas, aunque estas resultan a ser más complicadas (figura 3.1).



3.3 Dependencia funcional e independencia

La relación entre las variables X e Y, parte del objetivo de este capítulo y en general de un número importante de los estudios de las Ciencias Sociales, puede ser más o menos acentuada, pudiendo llegar ésta desde la dependencia total o dependencia funcional hasta la independencia.

3.3.1 Dependencia funcional

La dependencia funcional, que nos refleja cualquier fórmula matemática o física, es a la que estamos normalmente más habituados. Al principio del capítulo consideramos un ejemplo en el que sobre una población de alumnos definíamos las variables

$$\begin{aligned} X &\equiv \text{altura medida en centímetros,} \\ Y &\equiv \text{altura medida en metros,} \end{aligned}$$

Al tomar a uno de los alumnos, hasta que no se realice una medida sobre el mismo, no tendremos claro cual será su altura. Podemos tener cierta intuición sobre qué valor es más probable que tome (alrededor de la media, con cierta dispersión). Sin embargo, si la medida X ha sido realizada, no es necesario practicar la de Y, pues la relación entre ambas es exacta (dependencia funcional):

$$Y = X/100$$

Ello puede describirse como que conocido el valor $X=x_i$, la distribución de $Y|X=x_i$ sólo toma un valor con frecuencia del 100%. Esto se traduce en una tabla bidimensional de X e Y, del siguiente modo: La variable Y depende funcionalmente de la variable X si para cada fila $X=x_i$, existe un único j tal que $n_{ij} \neq 0$. Análogamente, tenemos dependencia funcional de X con respecto a Y haciendo el razonamiento simétrico, pero por columnas, es decir, X depende funcionalmente de la variable Y si para cada columna $Y=y_j$, existe un único i tal que $n_{ij} \neq 0$. Es claro que si la dependencia funcional es recíproca, la tabla es necesariamente cuadrada ($k=p$).

Ejemplo

Consideramos una población formada por 12 individuos, donde hay 3 franceses, 7 argentinos y 3 guineanos. Definimos las variables:

- X = Continente de nacimiento** ~ {Europa, América, África}
- Y = Nacionalidad** ~ {Francés, Guineano, Argentino}
- Z = Hablar español** ~ {Si, No}

Entonces, sobre esta población, podemos construir las siguientes tablas:

Z	Si	No	
X			
Europa	0	3	3
América	7	0	7
África	2	0	2
	9	3	12

Y	Francés	Guineano	Argentino	
X				
Europa	3	0	0	3
América	0	0	7	7
África	0	2	0	2
	3	2	7	12

y nos damos cuenta de que, según la definición Z depende funcionalmente de X.

- X no depende funcionalmente de Z.
- X e Y depende funcionalmente la una de la otra de modo recíproco.

3.3.2 Independencia

Hemos visto que la dependencia funcional implica una estructura muy particular de la tabla bidimensional, en la que en todas las filas (o en todas las columnas) existe un único elemento no nulo. Existe un concepto que de algún modo es el opuesto a la dependencia funcional, que es el de independencia. Se puede expresar de muchas maneras el concepto de independencia, y va a implicar de

nuevo una estructura muy particular de la tabla bidimensional, en el que todas las filas y todas las columnas van a ser proporcionales entre sí.

Para enunciar lo que es la independencia de dos variables vamos a basarnos en el siguiente razonamiento: Si la variable Y es independiente de X, lo lógico es que la distribución de frecuencias

relativas condicionadas $Y|_{x_1}$ sea la misma que la de $Y|_{x_2}, \dots, Y|_{x_h}$. Esto se puede escribir diciendo que

$$\forall j = 1, \dots, p, \text{ se tiene que } f_j^1 = \dots = f_j^i = \dots = f_j^h = f_{\cdot j}$$

Pues bien, diremos que la variable Y es independiente de X si la relación (3.3) es verificada. Hay otras formas equivalentes de enunciar la independencia: Cada una de las siguientes relaciones expresa por sí sola la condición de independencia:

3.3.2.1 Proposición (Independencia en tablas de doble entrada)

Cada una de las siguientes relaciones expresa por sí sola la condición de independencia entre las variables X e Y

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

$$\frac{n_{1j}}{n_{1\cdot}} = \frac{n_{2j}}{n_{2\cdot}} = \dots = \frac{n_{ij}}{n_{i\cdot}} = \dots = \frac{n_{kj}}{n_{k\cdot}} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$$

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}}$$

3.3.2.2 Observación

Obsérvese que la relación (3.4) (o bien la (3.5)) implica que la independencia es siempre recíproca, es decir, si X es independiente de Y, entonces Y es independiente de X.

3.3.2.3 Ejemplo

Si tenemos dos variables que son

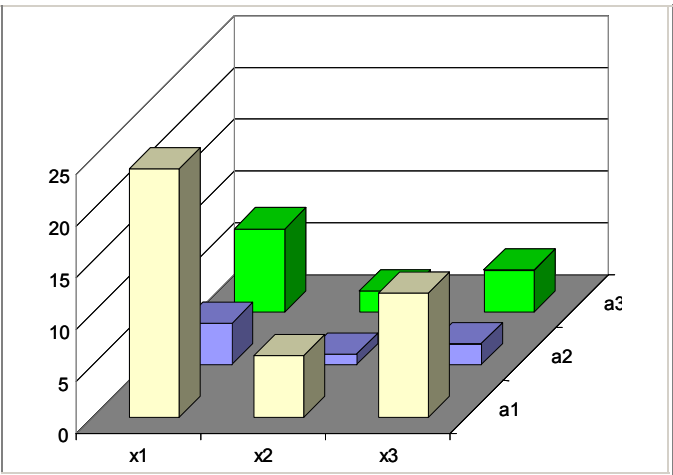
X ≡ número de claridifurcios melatonómicos

Y ≡ coeficiente de saturación de ciclopondrinas

y están distribuidas en una tabla del modo:

Y	1 ∈ (0, 2]	3 ∈ (2, 4]	5 ∈ (4, 6]		Figura: Cuando las variables son independientes, las diferencias entre las filas (o columnas) pueden entenderse como cambios de escala.
X					
0	24	4	8	36	
1	6	1	2	9	

2	12	2	4	18
	42	7	14	63



podemos decir que ambas variables son independientes. Obsérvese la proporcionalidad existente entre todas las filas de la tabla (incluidas la marginal) (figura 3.2). Lo mismo ocurre entre las columnas.

Bibliografía

1. ARMITAGE P, G. BERRY G. (1992) Estadística para la Investigación Biomédica. Doyma, Barcelona, 1992.
2. HAMILTON LC. (1992) Modern Data Analysis. Brooks/Cole Publishing Company, Pacific Grove, 1990.
3. MARTÍN ANDRÉS J.D. LUNA DEL CASTILLO (1992) Bioestadística para las Ciencias de la salud. Norma, Granada, 1994.
4. PEÑA SÁNCHEZ DE RIVERA D. (1992) Estadística: Modelos y Métodos, 1. Alianza Universidad Textos, Madrid, 1994.
5. RIVAS MOYA T, MATEO MA, RÍUS DÍAZ F, RUIZ M. (1992) Estadística Aplicada a las Ciencias Sociales: Teoría y Ejercicios (EAC). Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1991.
6. RUBIO CALVO E, T. MARTÍNEZ TERRER T Y OTROS (1992) Bioestadística. Colección Monografías Didácticas, Universidad de Zaragoza, Zaragoza, 1992.
7. SÁNCHEZ FONT E, RÍUS DÍAZ F. (1992) Guía para la Asignatura de Bioestadística (EAC). Secretariado de Publicaciones de la Universidad de Málaga, Málaga, 1990.

Bibliografía Adicional

1. Anderson GL, Prentice RL (1999) Individually randomized intervention trials for disease prevention and control. *Statistical Methods in Medical Research*, 1999; 8: 287±309
2. Armstrong BK, White E, Saracci R. (1994) Principles of Exposure Measurement in Epidemiology. Monographs in Epidemiology and Biostatistics. Volumen 21. Oxford University Press, Oxford.
3. Arto KA, Hawk DL. (1999) Industry Models of Risk Management and their Future. Proceedings of the 30th Annual Project Management Institute 1999 Seminars & Symposium Philadelphia, Pennsylvania, USA
4. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, Lipsitz S, Ryan L. (2000) Analysis of dichotomous outcome data for community intervention studies. *Statistical Methods in Medical Research*, 2000; 9:135±159
5. Chambers J et al. (1983) Graphical methods for data analysis. Wadsworth advanced books and software, Monterey, California, 1983
6. Cleveland W. (1985) The elements of graphing data. Wadsworth advanced books and software, Monterey, California, 1985
7. Daniel W. (1996) Bioestadística. Base para el análisis de las ciencias de la salud. Uteha, Noriega Editores. México, 1996
8. DeLucca P, Raghavarao D, Altan S. (1999) Effect of investigator bias on the power and level of the two-sample Z-test. *Journal of Biopharmaceutical Statistics*, 1999; 9(2), 279–288
9. Desrosières, A. (1996) Reflejar o instituir: la invención de los indicadores estadísticos. *Metodológica*, 1996; 4:41-56
10. Hersh AL, Black WC, Tosteson AN. (1999) Estimating the population impact of an intervention: a decision-analytic approach. *Statistical Methods in Medical Research*, 1999; 8: 311±330
11. Kieser M, Hauschke D. (1999) Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics*, 1999; 9(4), 641–650
12. Levy PS, Stolte K. (2000) Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Statistical Methods in Medical Research*, 2000; 9: 41±55
13. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, Meulen JHP, Bossuyt PMM. (1999) Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests. *JAMA*, 1999; 282(11): 1061-1066
14. Lin SCC. (1999) Some results on combinations of two binary screening tests. *Journal of Biopharmaceutical Statistics*, 1999; 9(1):81–88
15. Norman G, Sreiner D. (1996) Bioestadística. Mosby/Doyma libros. España, 1996

16. Obuchowski NA. (1998) Sample size calculations in studies of test accuracy. Department of Biostatistics and Epidemiology, The Cleveland Clinic. Foundation, Cleveland, Ohio, USA. *Statistical Methods in Medical Research*, 1998; 7:371±392
17. Patel HI, Rowe E. (1999) Sample size for comparing linear growth curves. *Journal of Biopharmaceutical Statistics*, 1999; 9(2), 339–350
18. Richard Simon R. (1999) The role of statisticians in intervention trials. National Cancer Institute, Bethesda, Maryland, USA. *Statistical Methods in Medical Research*, 1999; 8:281±286
19. Silva LC. (1997) *Cultura estadística e investigación científica en el campo de la salud. Una mirada crítica.* Díaz de Santos, España, 1997
20. Silva LC. (2000) *Diseño razonado de muestras y captación de datos para la Investigación Sanitaria.* Díaz de Santos, España, 2000.
21. Velleman PF, Wilkinson, L. (1993) Nominal, ordinal, interval, and ratio typologies are Misleading. *The American Statistician*, 1993; 47(1):65-72