Probabilidad y Estadística

Ángel Corberán
Francisco Montes

Departament d'Estadística i Investigació Operativa
Universitat de València

Departament d'Estadística i Investigació Operativa
Universitat de València
46100-Burjassot
Spain

TEMA 1.- DESCRIPCIÓN DE DATOS

1. Introducción

La Tabla 1 recoge, parcialmente, el resultado de una encuesta[1] a la que fueron sometidas 250 personas con edad igual o superior a 15 años, tabla que aparece completa en el Anexo I. Las columnas de las respuestas están encabezadas por nombres, abreviados en algunos casos, que hacen referencia a la pregunta formulada. Su significado y el de la codificación correspondiente es el siguiente:

- La primera columna indica el número de caso
- **Sexo:** indica el sexo del entrevistado, **v** = varón, **m** = mujer
- **E_civil:** indica el estado civil, **1** = casado/a, **2** = soltero/a, **3** = viudo/a, **4** = div/sep
- **Edad:** edad expresada en años
- **Niv_ed:** nivel de educación, **1** = analfabeto/a, **2** = sin estudios, **3** = est. primarios, **4** = BUP o similares, **5** = est. universitarios
- **Peso:** peso expresado en kilogramos
- **Altura:** altura expresada en centímetros
- **Jueces:** opinión sobre los jueces "estrella", **1** = buena, **2** = indiferente, **3** = mala, **4** =ns/nc
- **E_penal:** opinión sobre el adelanto de la edad penal, **1** = a favor, **2** = en contra, **3** =ns/nc
- **35horas:** opinión sobre la semana laboral de 35 horas **1** = a favor, **2** = en contra, **3** =ns/nc
- **C_alcohol:** consumo medio diario de alcohol medido en el equivalente a vasos de vino de 200cc, la escala va de 1 a 5, indicando esta última cifra 5 o más vasos diarios

	SEXO	E_CIVIL	EDAD	NIV_ED	PESO	ALTURA	JUECES	E_PENAL	35HORAS	C_ALCOHOL
01	v	1	63	3	80,30	190	3	1	1	3
02	v	1	79	4	56,16	155	1	3	1	2
03	m	1	52	3	64,37	151	3	3	1	2
04	m	3	41	3	63,02	146	3	2	2	2
05	v	2	18	4	75,50	164	4	2	1	3
06	m	2	68	3	35,00	136	4	3	2	2
07	v	2	35	2	62,79	145	3	1	1	2
08	m	2	46	2	78,92	190	1	2	1	3
09	m	2	20	3	58,27	171	3	2	1	0
10	v	1	61	4	52,17	159	3	2	2	2
11	m	1	69	3	70,82	169	3	1	2	2
12	m	2	50	3	41,10	167	4	2	1	3
13	m	1	67	2	49,46	171	3	2	3	1

Tabla 1.- **Reproducción parcial de las 250 observaciones del Anexo I**

Interpretar los datos que aparecen en la tabla presenta dificultades incluso para las personas con conocimientos de Estadística y, desde luego, prácticamente imposible para lo que podríamos denominar *gran público*. No por casualidad cuando se ofrece información de este tipo aparece resumida y transformada para hacerla fácilmente comprensible, resumen que pretende llamar nuestra atención sobre los aspectos más relevantes de los datos y que para conseguirlo utiliza las herramientas propias de la Estadística Descriptiva o Descripción de datos, a saber:

- distribuciones de frecuencia,
- gráficos,
- medidas de posición o centrales, y
- medidas de dispersión.

Antes de comenzar el resumen de los datos de nuestra tabla, introduciremos el lenguaje y las definiciones que nos permitan hacerlo.

[1] La encuesta es ficticia y las respuestas que en ella figuran han sido simuladas. Se trata tan solo de un ejemplo elaborado a los efectos de presentación y desarrollo del temario.

2. Muestra y variables

Los datos recogidos en la tabla reciben el nombre de **muestra**, que a su vez está constituida por las **observaciones muestrales** a cuyo número denominaremos **tamaño muestral**. En cada observación hay una o varias **variables observadas**. En el caso de nuestra tabla tenemos:

- una **muestra** que contiene las respuestas a determinada encuesta,
- el **tamaño muestral** es de 250 observaciones
- cada **observación muestral** se corresponde con las respuestas a la encuesta de una persona con edad igual o superior a 15 años,
- las **variables observadas** son: *sexo, estado civil, nivel de educación, peso, altura, opinión sobre los jueces, opinión sobre edad penal, opinión sobre semana laboral de 35 horas* y *consumo medio diario de alcohol.*

Las variables, lógicamente, han de centrar nuestra atención prioritariamente, razón por la cual conviene establecer una clasificación de las mismas:

$$\textit{tipos de variables}\begin{cases}\textit{cualitativas}\begin{cases}\textit{ordinales}\\ \textit{no ordinales}\end{cases}\\ \textit{cuantitativas}\begin{cases}\textit{discretas}\\ \textit{continuas}\end{cases}\end{cases}$$

variables cualitativas: son variables que describen categorías, razón por la cual se las denomina también **categóricas**. Cuando las categorías admiten algún tipo de ordenación se las denomina **ordinales** (por ejemplo, la variable *nivel de educación* de la tabla) y **no ordinales** en caso contrario (por ejemplo, las variables *sexo, estado civil, opinión sobre los jueces, opinión sobre edad penal, opinión sobre semana laboral de 35 horas)*

variables cuantitativas: son variables que expresan valores numéricos, **discretas** o **continuas** según la naturaleza de la observación. En la tabla, *consumo medio diario de alcohol* es un ejemplo de las primeras y *peso, altura* son ejemplos de las segundas.

La frontera entre variables discretas y continuas es en ocasiones difusa debido a la acción discretizadora que todo proceso de medida comporta. En efecto, si observamos la variable *edad* en la tabla nadie pondrá en duda su carácter continuo pues mide el *tiempo* transcurrido desde el nacimiento de una persona, pero, en general, las fracciones de año son irrelevantes razón por la cual viene medida en años y aparece expresada mediante valores enteros positivos.

3. Distribuciones de frecuencias

Una primera descripción resumida de los datos puede llevarse a cabo mediante la distribución de frecuencias de cada una de las variables. Como luego pondremos de manifiesto, el tipo de variables es determinante a la hora de analizar los datos con esta herramienta. Para variables categóricas y discretas con un rango pequeño de valores utilizaremos distribuciones de frecuencias no agrupadas de las que nos ocupamos a continuación:

Frecuencias no agrupadas Se trata simplemente de obtener y representar gráficamente el número de ocurrencias (**frecuencia absoluta)** de las distintas categorías o valores de la variable. En ocasiones es conveniente utilizar la **frecuencia relativa**, definida como:

$$frecuencia\ relativa = \frac{frecuencia}{n},$$

donde n es el tamaño muestral. La frecuencia relativa se suele expresar también en porcentaje. Obtengamos la distribución de frecuencias asociada a alguna de las variables de la tabla.

Estado civil	Frec	Frec Rel
casado/a	131	0.524
soltero/a	98	0.392
viudo/a	16	0.064
div/sep	5	0.020
Total	250	1.00

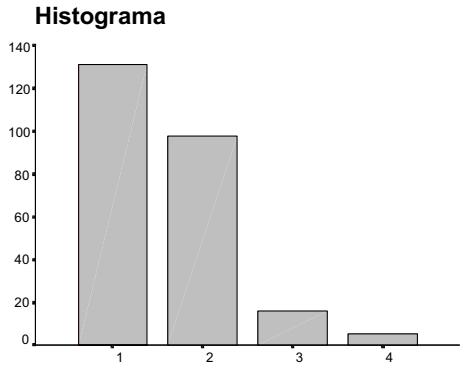


Figura 1.- **Tabla de frecuencias e Histograma de E_CIVIL**

Opinión	Frec	Frec Rel
buena	67	0.268
indiferente	49	0.196
mala	112	0.448
ns/nc	22	0.088
Total	250	1.00

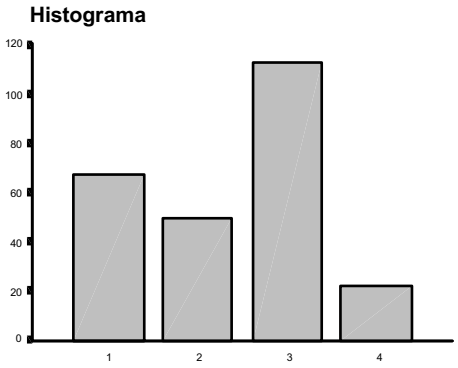


Figura 2.- **Tabla de frecuencias e Histograma de JUECES**

Vasos	Frec	Frec Rel
0	22	0.088
1	50	0.200
2	90	0.360
3	49	0.196
4	26	0.104
5	13	0.052
Total	250	1.00

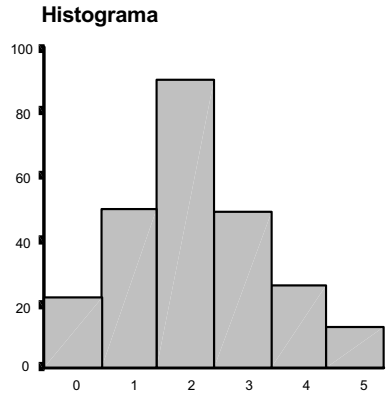


Figura 3.- **Tabla de frecuencias e Histograma de C_ALCOHOL**

La representación gráfica de las frecuencias, en los casos de variables categóricas o de variables discretas con pocos valores, puede también llevarse a cabo mediante **Diagramas de Sectores**, en los que cada valor o categoría de la variable se representa mediante un sector circular con área proporcional a su frecuencia. Las figura 4 y 5 son una muestra de estos diagramas para las variables *nivel estudios* y *35horas.*

Estudios	**Frec**	**Frec Rel**
analf.	7	0.028
sin est.	57	0.228
primarios	112	0.448
BUP o sim.	61	0.244
univers.	13	0.052
Total	250	1.00

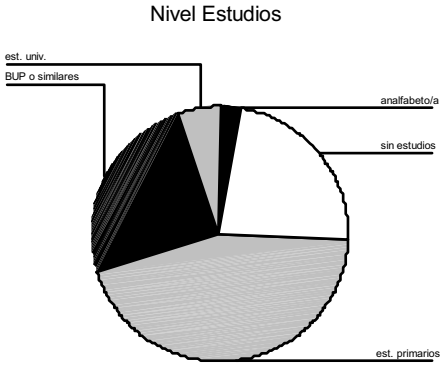


Figura 4.- **Tabla de frecuencias y Diagrama de Sectores de NIV_ED**

35 horas	**Frec**	**Frec Rel**
a favor	131	0.524
en contra	96	0.384
ns/nc	23	0.092
Total	250	1.00

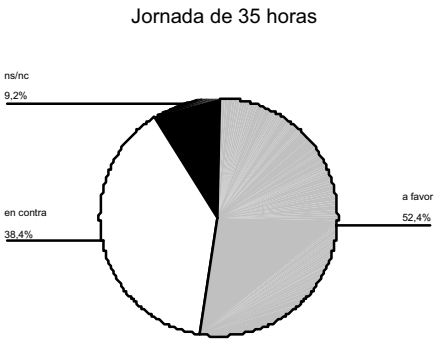


Figura 5.- **Tabla de frecuencias y Diagrama de Sectores de 35HORAS**

Frecuencias agrupadas Si pretendemos resumir la información de las variables *edad*, *altura* y *peso* tal como lo hemos hecho en las anteriores, es decir, considerando cada valor como una categoría obtendremos una tabla de frecuencias y un histograma que, al no condensar la información, nos servirán de poca ayuda. El motivo está en el carácter continuo de la variable.
El problema se resuelve agrupando los valores de la variable en **clases** y obteniendo la distribución de frecuencias para dichas clases.
Las clases son intervalos y están delimitadas por los **límites de clase**, y deben constituir una partición del conjunto de valores que toma la variable, es decir, las clases no se solapan y no deben excluir ningún valor de la variable, lo que permite clasificar a cualquier valor en una y solo una de las clases establecidas. La distancia entre los límites de la clase es la **amplitud de la clase**.
En la gráfica siguiente aparece la distribución de frecuencias de la variable *edad* que ha sido agrupada en los intervalos que se indica en la tabla, a saber, 8 clases de longitud 10, donde la clase i-ésima es el intervalo $[x_i, x_{i+1}[$, que al estar abierto en su límite superior no se solapa con la clase siguiente.

Edad	Frec	Frec Rel
15-25	48	0.192
25-35	43	0.172
35-45	50	0.200
45-55	34	0.136
55-65	23	0.092
65-75	30	0.120
75-85	16	0.064
85-95	6	0.024
Total	250	1.00

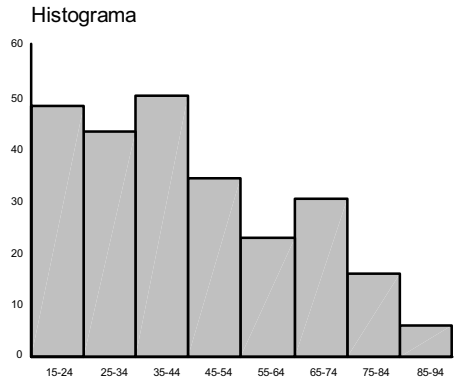
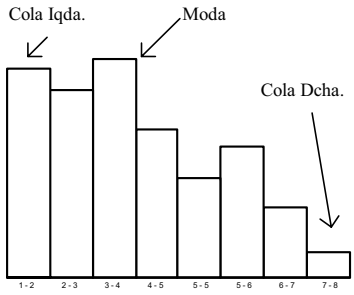


Figura 6.- **Tabla de frecuencias e Histograma de EDAD**

¿Qué información de interés nos proporciona el histograma anterior? Aunque más tarde estudiaremos con mayor detalle este problema, adelantemos ya algunos aspectos relevantes de la forma de la distribución de frecuencias. El pico, que representa la mayor frecuencia es la **moda**, valor alrededor del cual se distribuyen los valores que toma la variable, cuyas frecuencias van disminuyendo a derecha e izquierda para formar en los extremos las llamadas colas de la distribución. En nuestro caso, la **cola izqda.** es más *pesada* que la **derecha**, indicando con ello que hay mayor presencia de edades inferiores que de superiores y dando lugar a una distribución sin **simetría** y **sesgada** a la izquierda.



Número de clases a establecer. La pregunta que surge al observar la distribución de frecuencias anterior es ¿por qué 8 clases y no 14? No es difícil imaginar que un número de clases distinto producirá una gráfica de aspecto diferente, como puede observarse en los histogramas que aparecen a continuación; en ellos la variable edad ha sido representada con 3 y 30 clases, respectivamente.

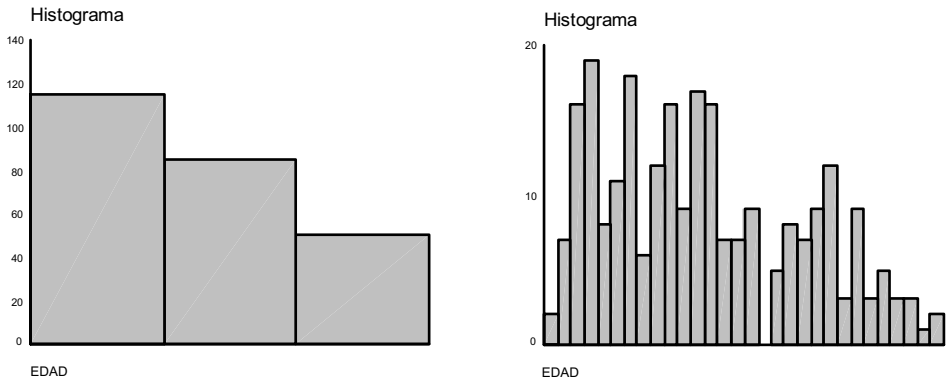


Figura 7.- **Histogramas de barras para EDAD con 3 y 30 clases**

No debemos olvidar que el objetivo de cualquier agrupación de datos es destacar los rasgos esenciales y eliminar los detalles irrelevantes, aún cuando esto se haga a expensas de perder una información que no consideramos esencial, de ahí la importancia de elegir adecuadamente el número y amplitud de las clases. Las siguientes recomendaciones pueden ayudarnos, aunque puede ser conveniente llevar a cabo distintas elecciones y comparar los resultados:

- si el tamaño de la muestra es $n \leq 50$, un número de clases entre 5 y 15 suele ser apropiado; para muestras mayores este número puede superar las 20 clases,
- el rango de la variable, que es *rango = valor mayor - valor menor*, y la amplitud que deseamos para cada clase nos permitirán determinar su número. Por ejemplo, para la tabla de frecuencias e histograma de la variable *edad* que hemos representado en la figura 6, hemos calculado su rango = 92-15 = 77 y como deseábamos una amplitud de 10 años para cada clase, hemos obtenido un número de 7.7, que lógicamente se ha redondeado a 8, lo que supone que la última clase cubre el intervalo [85,95[,

Clases con amplitudes distintas Los histogramas que hemos utilizados hasta ahora provienen de distribuciones de frecuencias agrupadas cuyas clases tienen todas igual amplitud, razón por la cual su **altura** es directamente proporcional a su frecuencia.
Cuando las frecuencias de clases contiguas son bajas pueden agruparse en clases mayores cuya frecuencia será la suma de las frecuencias de las clases que constituyen la nueva clase. Por ejemplo, lo datos de la Tabla 2 son una muestra de 30 valores de la variable *peso*, extraídos de entre los 250 que constituyen los datos originales. La tabla de frecuencias muestra que la segunda clase, [35,45[, tiene una frecuencia 0.

81,72	52,44	69,24	58,34	81,43	52,35	28,60	92,78	87,82	59,44
86,39	68,26	57,29	83,62	26,14	68,47	56,00	96,97	57,79	65,10
78,37	56,74	45,41	65,85	48,95	81,84	74,82	91,93	71,48	68,34

Tabla 2.- **30 observaciones del peso**

Peso	**Frec**	**Frec Rel**
25-35	2	0.066
35-45	0	0.000
45-55	4	0.134
55-65	6	0.200
65-75	8	0.267
75-85	5	0.166
85-95	4	0.134
95-105	1	0.033
Total	30	1.00

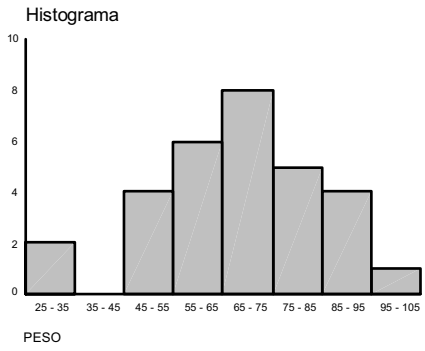


Figura 8.- **Tabla de frecuencias e Histograma de los 30 valores del *peso***

Puede pensarse en la conveniencia de agrupar las dos primeras clases en una sola, [25,45[, para conseguir una distribución de frecuencia más suavizada que evite la frecuencia 0. La consecuencia de esta agrupación es una distribución de frecuencias con clases de distinta amplitud, una de ellas el doble que las restantes, y debemos cambiar el método de representación del histograma para evitar distorsiones en su forma. En efecto, si, como hasta ahora, la altura de la barra correspondiente a cada clase es proporcional a su frecuencia, obtendremos el histograma B de la figura 9, que transmite visualmente la idea de una presencia de la primera clase mayor de la que en realidad le corresponde. Esto se evita haciendo que las **áreas** de las barras sean proporcionales a la frecuencia, como se ha hecho en el histograma A, lo que conduce en nuestro caso a una altura que es la mitad de la anterior puesto que la base del rectángulo es el doble.

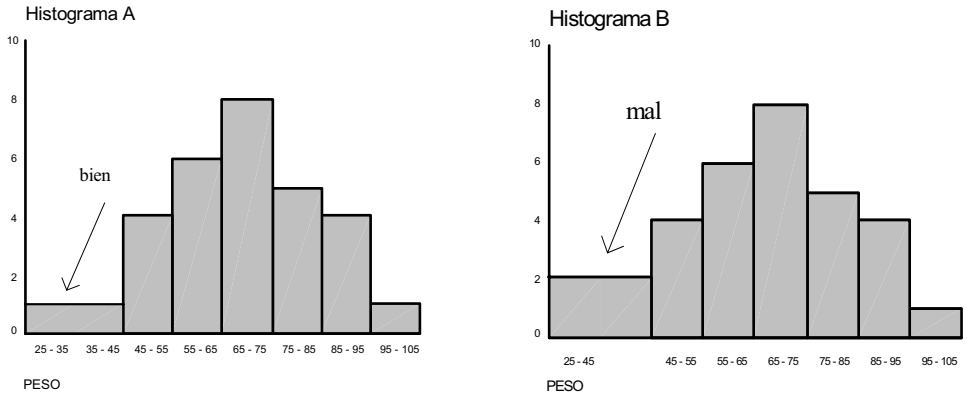


Figura 9.- **Histogramas de frecuencias para distribuciones con clases de distinta amplitud**

4. Medidas de Posición

Para las variables categóricas, las distribuciones de frecuencias y sus distintas representaciones gráficas nos proporcionan información concisa y completa, pero si las variables son cuantitativas es posible, y conveniente, completar aquella información con características numéricas asociadas a los datos. Estas características reciben el nombre de **estadísticos descriptivos** y los hay de dos tipos: de **posición** o **centrales** y de **dispersión**. Los primeros nos proporcionan información acerca de la posición de los datos si los representamos en una recta, mediante la obtención de lo que podríamos llamar *centro* de la distribución. Existen distintas formas de definir el centro de una distribución de datos, las más utilizadas son: *la media*, *la mediana*, *la moda y los percentiles*.

En adelante designaremos mediante las últimas letras mayúsculas del abecedario, **X**, **Y**, **Z**, …, a las variables observadas y con las minúsculas, **x**, **y**, **z**, …, las observaciones (datos), a las que cuando sea conveniente añadiremos un índice. Por ejemplo, si queremos designar las *n* observaciones de la variable **X** lo podemos hacer mediante $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$, …, $\mathbf{x_n}$.

La media Es sin duda la más conocida de las medidas de posición y es, sencillamente, la **media aritmética** de las observaciones correspondientes a la variable en estudio. Se le denomina **media muestral** y se le designa mediante el símbolo $\overline{\mathbf{x}}$. Su expresión es,

$$\overline{x} = \frac{suma\ de\ las\ x's}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Retomemos los datos de las 30 observaciones de pesos contenidos en la Tabla 2, para calcular su media

$$\overline{x} = \frac{81{,}72+52{,}44+69{,}24\ +\cdots+91{,}93+71{,}48+68{,}34}{30} = \frac{2013{,}92}{30} = 67{,}13\ kgs.$$

La mediana Es aquel valor que, al ordenar las observaciones de menor a mayor, ocupa el lugar central, dividiendo el conjunto de observaciones en partes iguales. Es decir, que deja a su derecha y a su izquierda el 50% de las observaciones. Si el tamaño de la muestra, *n*, es **impar**, necesariamente existe una observación que ocupa el lugar central, concretamente la que al ordenar las observaciones está en la posición (n+1)/2; si, por contra, *n* es **par**, son dos las observaciones que ocupan el lugar central, las que están en las posiciones n/2 y (n/2)+1, definiéndose entonces la mediana como el punto medio entre ambas observaciones. Veamos algunos ejemplos:

- **Ejemplo 1:** Si ordenamos los 30 valores del *peso* de la Tabla 2 tendremos:

  26,14 28,60 45,41 48,95 52,35 52,44 56,00 56,74 57,29 57,79
  58,34 59,44 65,10 65,85 **68,26 68,34** 68,47 69,24 71,48 74,82
  78,37 81,43 81,72 81,84 83,62 86,39 87,82 91,93 92,78 96,97

  y siendo *n=30* par, la mediana será el valor medio de los valores que ocupan las posiciones 15 y 16, que aparecen en negrita en la ordenación. Así pues,

$$\text{mediana} = \frac{68{,}26 + 68{,}34}{2} = 68{,}30 \ kgs.,$$

  valor que, como puede observarse, no coincide con el de la media antes calculada.

- **Ejemplo 2:** Las 13 primeras observaciones correspondientes al *consumo de alcohol* ordenadas de menor a mayor son: 0 1 2 2 2 2 **2** 2 2 3 3 3 3. La que ocupa la posición central, la séptima puesto que hay 13 valores, es la mediana y su valor es 2.

La moda Es aquel valor de la variable que tiene mayor frecuencia. En el caso de frecuencias agrupadas se toma la clase más frecuente como moda. Así, para la variable *consumo de alcohol* la moda es 2 (ver tabla de frecuencias de la Figura 3) y para la variable *edad* la moda es la clase 35-45 (ver Figura 6).

Los percentiles El percentil **p-ésimo** es aquel valor que verifica la condición de que un p% de las observaciones son menores o iguales que el. Así, el percentil 70-ésimo supone que el 70% de las observaciones son menores o iguales que el valor de dicho percentil.
La Tabla 3 nos muestra, ordenadas de izquierda a derecha y de arriba a abajo, las 250 observaciones correspondientes a la variable *altura*. La primera fila y la primera columna, en negrita, han sido añadidas para mejor localizar las posiciones de cada valor en la ordenación. Así, si queremos conocer el percentil 30-ésimo, tendremos en cuenta que el 30% de 250 es 75 y buscaremos el valor que ocupa esta posición en la tabla, el 162. El percentil 15-ésimo es 154 porque, aunque el 15% de 250 es 37.5, los valores correspondientes a las posiciones 37 y 38 son, ambos, 154. Si no hubiera sido así, hubiéramos tomado el valor correspondiente a la posición más cercana. De la misma manera calcularíamos el percentil 90 que es 190.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
0	109	123	126	132	135	136	136	138	141	141	142	142	143	144	145	145	146	147	147	147
20	148	148	148	149	149	150	150	150	151	151	151	151	152	152	153	153	**154**	154	155	155
40	155	156	157	157	157	157	157	158	158	158	159	159	159	159	159	159	160	160	160	160
60	160	160	160	161	161	161	161	161	162	162	162	162	162	162	**162**	163	164	164	164	164
80	164	164	164	165	165	165	165	165	165	165	165	166	166	166	167	167	167	167	167	167
100	167	168	168	168	168	168	168	169	169	169	169	169	169	169	169	169	170	170	170	170
120	170	170	170	171	171	171	171	171	171	171	172	172	172	172	172	172	173	173	173	173
140	173	173	173	173	174	174	174	174	174	174	174	175	175	175	175	175	175	176	176	176
160	176	176	177	177	177	177	177	177	177	178	178	178	178	178	178	179	179	179	179	180
180	180	180	181	181	181	182	182	182	182	182	182	182	182	183	183	183	183	184	184	184
200	184	185	185	185	185	185	185	186	186	186	186	187	187	187	187	187	187	187	189	189
220	189	189	189	189	**190**	190	190	190	190	190	191	192	192	192	192	194	195	195	197	198
240	200	200	201	202	202	203	207	215	218	218										

Tabla 3.- **Las 250 observaciones de la variable *altura* ordenadas**

Los percentiles 25, 50, y 75-ésimo reciben el nombre de **primer cuartil**, **segundo cuartil** y **tercer cuartil**, respectivamente. El nombre les viene de dividir las observaciones en cuartos. Observemos que según la definición que hemos dado para la mediana, ésta coincide con el percentil 50-ésimo o segundo cuartil.

5. Medidas de Dispersión

La medidas de posición nos dan una información incompleta, por parcial, acerca de las observaciones. En efecto, supongamos que las notas de Matemáticas de los estudiantes pertenecientes a dos clases distintas, clase I y clase II con 10 estudiantes cada una, son las siguientes:

clase I: 4, 3, 5, 6, 4, 5, 5, 7, 5, 6
clase II: 1, 4, 3, 5, 6, 8, 2, 7, 5, 9

en ambos casos la media, como puede comprobarse con facilidad, es 5, pero sus histogramas de frecuencias son muy distintos.

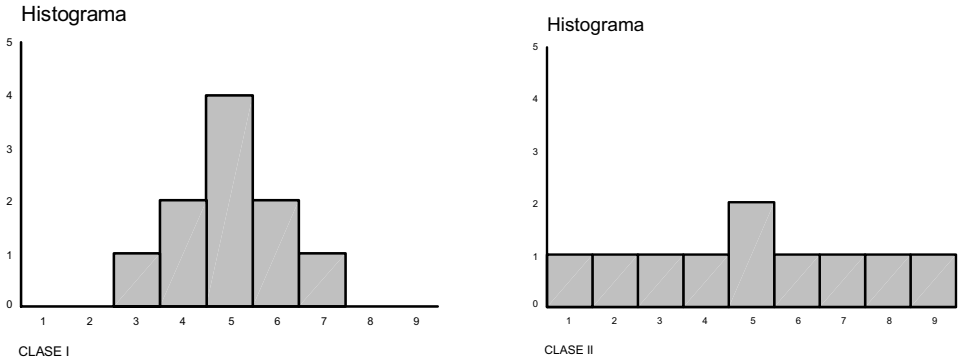


Figura 10.- **Histogramas de frecuencias para notas de las clases I y II**

La Figura 10 muestra que los valores se distribuyen simétricamente respecto de la nota 5, pero en la clase I existe una dispersión menor que en la clase II. ¿Cómo medir la distinta manera en que los valores se agrupan alrededor de la media? Las distintas **medidas de dispersión** proporcionan esta información. Al igual que ocurre para la posición, existen diversas formas de medir la dispersión, de entre ellas vamos a ocuparnos de las siguientes: *rango, desviación típica, varianza* y *rango intercuartílico*.

El rango Es la diferencia entre el máximo y el mínimo de las observaciones. Así, para los datos anteriores tendremos que rango de las notas en la clase I vale **4** y el rango en la clase II vale **8**, denotando la mayor dispersión de la variable en el segundo grupo de observaciones.

La varianza y la desviación típica Puesto que se trata de medir cómo se agrupan los valores alrededor de la media, podríamos utilizar como criterio las desviaciones de dichos valores respecto de aquella, es decir, la diferencias entre la media y los distintos valores y más concretamente la media de ellas. Aunque a primera vista la sugerencia pueda ser buena, vamos a aplicarla a los valores de las notas de clase para evidenciar el inconveniente insalvable que una medida de este tipo tiene.
En el cuadro aparecen las notas de cada clase y en columnas sucesivas sus desviaciones respecto de la media y el cuadrado de estas desviaciones, al que más tarde aludiremos. Al tratar de obtener la media de las diferencias, que recordemos es la suma de todas ellas dividida por su número, nos encontramos que dicha media será **0** en ambos casos, porque existiendo desviaciones positivas y negativas, unas anulan los efectos de las otras. En realidad eso nos ocurrirá con cualquier otro conjunto de datos, porque puede demostrarse que esa es una propiedad que tienen las desviaciones respecto de la media.

CLASE I			CLASE II		
nota	$\mathbf{d_i = x_i - \bar{x}}$	$\boldsymbol{d_i^2}$	**nota**	$\mathbf{d_i = x_i - \bar{x}}$	$\boldsymbol{d_i^2}$
4	1	1	1	4	16
3	2	4	4	1	1
5	0	0	3	2	4
6	-1	1	5	0	0
4	1	1	6	-1	1
5	0	0	8	-3	9
5	0	0	2	3	9
7	-2	4	7	-2	4
5	0	0	5	0	0
6	-1	1	9	-4	16
Suma	**0**	**12**	**Suma**	**0**	**60**

Tabla 4.- **Desviaciones respecto de la media y sus cuadrados para las notas de las clase I y II**

Puesto que el uso de las desviaciones respecto de la media parece razonable, ¿cómo soslayar el problema? Una manera sencilla de hacerlo es utilizar, no las desviaciones, sino sus cuadrados. Al ser estos cantidades positivas, su suma nunca podrá ser cero. Así, la media de los cuadrados de las desviaciones parece una medida adecuada, pero, por razones técnicas que están fuera del alcance y objetivos de este curso, la utilizaremos con una ligera modificación: en lugar de dividir por *n*, como se hace habitualmente para calcular una media, dividiremos por *n-1*. De acuerdo con esto, **la varianza** de un conjunto de observaciones se define mediante la fórmula:

$$s^2 = \frac{\textit{suma del cuadrado de las desviaciones}}{n-1} = \frac{\sum_{i=1}^{n} (\bar{x} - x_i)^2}{n-1}.$$

La **desviación típica** se define como la raíz cuadrada de la varianza y la designamos por *s*.
Para el caso de las clases I y II las sumas de los cuadrados de las desviaciones aparecen en la Tabla 4, sus varianzas y desviaciones típicas son:

clase I $\quad s^2 = \frac{12}{9} = 1{,}33 \qquad s = \sqrt{1{,}33} = 1{,}15$

clase II $\quad s^2 = \frac{60}{9} = 6{,}66 \qquad s = \sqrt{6{,}66} = 2{,}58$

que ponen de manifiesto la diferente distribución de los valores en un caso y otro. Para los 30 valores del *peso* de la Tabla 2,

peso $\quad s^2 = \frac{9040{,}76}{29} = 311{,}75 \text{ kg}^2 \qquad s = \sqrt{311.75} = 17{,}65 \text{ kg}.$

Obsérvese que las unidades de la varianza son el cuadrado de las unidades en las que venga expresada la variable, sin embargo la desviación no cambia de unidades.
Señalemos por último que si el tamaño de la muestra es grande, la diferencia entre dividir por *n* o por *n-1* es inapreciable y la varianza coincide, prácticamente, con la media de los cuadrados de las desviaciones.
Porcentajes típicos La desviación típica tiene una propiedad interesante, para distribuciones de frecuencias con una sola moda, de apariencia simétrica y con colas ni demasiado largas ni demasiado cortas, se suele verificar:

- aproximadamente el 68% de las observaciones distan como mucho **una** desviación típica de la media
- aproximadamente el 95% de las observaciones distan como mucho **dos** desviaciones típicas de la media
- aproximadamente más del 99% de las observaciones distan como mucho **tres** desviaciones típicas de la media

El rango intercuartílico Se define como la diferencia entre el tercer y el primer cuartil, IQR = Q_3 - Q_1. Directamente relacionado con él se define el **intervalo intercuartílico**, que es el intervalo definido por los cuartiles primero y tercero, [Q_1 , Q_3], cuya longitud es, precisamente, IQR. Contiene el 50% de las observaciones centrales. Para las 250 observaciones correspondientes a la *altura* estas medidas valen:

altura Q_1 = 160 cm. Q_3 = 182 cm. IQR = 22 cm.

El coeficiente de variación Aún cuando no se trata, estrictamente, de una medida de dispersión este es el momento de definir esta nueva característica asociada a las observaciones. Para comprender mejor su interés tratemos de responder a la pregunta, ¿dónde hay mayor dispersión, en las observaciones del peso o en las notas de la clase I? La pregunta tiene difícil respuesta si, por ejemplo, pretendemos comparar directamente las correspondientes desviaciones típicas. En efecto, la del *peso* es mucho mayor que la de las *notas*, pero a nadie se le escapa que la magnitud de aquel es mucho mayor que las de éstas y, además, se trata de unidades diferentes, kilogramos en un caso y puntuación en el otro. Para resolver el problema se define el **coeficiente de variación** como el cociente entre la desviación típica y la media multiplicado por 100,

$$CV = \frac{s}{\bar{x}} 100 ,$$

que expresa la desviación típica en porcentaje de la media y que al no tener unidades permite comparaciones entre observaciones de distinta naturaleza. Volviendo a la pregunta inicial, para el peso, CV_{peso} = 100×(17,65/67,13) = 26,29%, y para las notas, $Cv_{notas\ I}$ = 100×(1,15/5) = 23%, lo que nos dice que en términos de porcentaje de sus medias, ambas distribuciones tienen dispersiones muy parecidas.

6. Transformación de una variable: tipificación

En ocasiones puede ser interesante transformar los valores observados mediante cambios sencillos. Por ejemplo, multiplicarlos por una constante y/o sumarles alguna cantidad fija. Una transformación de este tipo se denomina *lineal* y se expresa mediante la fórmula:

$$Y = aX + b ,$$

donde **Y** es la nueva variable que resulta de multiplicar la variable original, **X**, por a y añadirle b al resultado. Cuando $b = 0$, la transformación recibe el nombre de **homotecia** o **cambio de escala**, si $a = 1$ y $b \neq 0$, recibe el nombre de **traslación**. Por ejemplo, si decidimos cambiar la escala en las alturas observadas y expresarlas en metros, $Y = X/100$, con $a = 1/100$, puesto que 1m = 100cms.
¿Cómo afecta una transformación lineal a las media, la varianza y la desviación típica? Con relativa sencillez puede comprobarse que éstas se ven afectadas de la manera que indica en el cuadro:

	X	**Y**
media	$\bar{x}$	$\bar{y} = a\bar{x} + b$
desviación típica	s_X	$s_Y = as_X$
varianza	s_X^2	$s_Y^2 = a^2 s_X^2$

Por ejemplo, si decidiéramos expresar los pesos en gramos, la transformación sería de la forma $Y=1000X$, y tendríamos $\bar{y}$ = 67130 gr., s_Y = 17656 gr. y s_Y^2 = 311750517 gr^2.

Finalicemos con una transformación que tiene nombre propio y que es muy utilizada en estadística, se trata de la **tipificación de una variable**. Es una transformación lineal en la que $a = 1/s_X$ y $b = \bar{x}/s_X$ y que consiste en:

$$Y = \frac{X - \bar{x}}{s_X},$$

es decir, restarle a cada valor la media y dividirlo luego por la desviación típica. Si tenemos en cuenta el efecto de la transformación descrito en el cuadro anterior, $\bar{y} = 0$, $s_Y = 1$ y $s_Y^2 = 1$, y a la nueva variable se la conoce con el nombre de **variable tipificada**. Obsérvese que cualquiera que sea **X** inicialmente, la variable tipificada correspondiente tiene siempre media 0 y varianza y desviación típica 1.
Como comentario final, que se deja a la comprobación del lector, el *coeficiente de variación* no se altera cuando se lleva a cabo un cambio de escala.

7. Muestra y Población: Inferencia Estadística

Solo en contadas ocasiones nuestro interés al analizar un conjunto de observaciones se limita a una simple descripción de las mismas mediante los distintos métodos explicados en párrafos anteriores. Casi siempre la descripción constituye un primer paso para conocer aquello que el conjunto de observaciones representa y que no nos es accesible.

Recordemos que al comenzar el tema aludíamos al conjunto de datos del Anexo I como a una **muestra** de personas con edad igual o superior a los 15 años. La pregunta que surge de inmediato es, *¿de dónde proviene esa muestra?* Para responder debemos introducir el concepto de **población**, que en nuestro caso podríamos definir como el total de individuos con edad igual o superior a 15 años. La pretensión habitual de los investigadores es conocer las características de la población a partir de lo observado en la muestra, en aquellas ocasiones en las que el estudio exhaustivo de la población es imposible en la práctica.

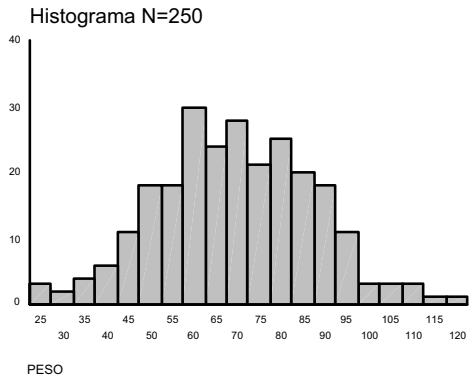
¿A qué nos referimos cuando hablamos de conocer la población a través de la muestra? Nos referimos a bajo qué condiciones, si la media de las 250 *alturas* de la muestra vale 170.68 cm., este valor puede tomarse como la altura media de las personas con edad igual o superior a 15 años. El proceso que estudia la manera de extraer conclusiones acerca de una población, partiendo de las observaciones contenidas en una muestra de aquella, se denomina **Inferencia Estadística**, y de él nos ocuparemos en temas posteriores. Podemos sin embargo adelantar la condición fundamental que toda muestra debe cumplir respecto de la población que pretende representar: la muestra debe ser *representativa*. Ello significa que ha de haber sido obtenida de tal manera que reproduzca los rasgos de aquella. Esto lo entenderemos mejor mediante algunos ejemplos de muestras que no serían representativas de las alturas de nuestra población original:

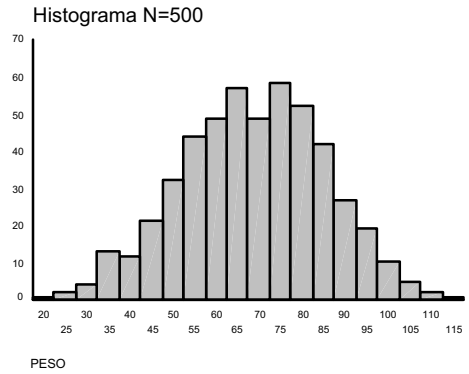
- que la muestra hubiera sido elegida sólo entre los hombres, en este caso podríamos extender nuestras conclusiones al conjunto de los hombres con edad igual o mayor a 15 años, pero de ninguna manera a todas las personas que cumplen la condición de edad,
- que la muestra hubiera sido elegida entre las personas con rentas altas, porque, si admitimos que mayor renta implica mejores condiciones de vida y alimentación, el resultado podría ofrecernos una altura media superior a la real.

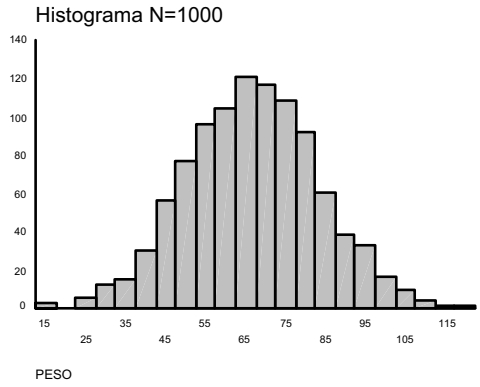
Bajo el nombre de **técnicas de muestreo** se conocen los distintos procedimientos que garantizan la representatividad de una muestra y estudian cómo el tamaño de la muestra influye en la calidad de nuestras conclusiones. Su importancia es obvia y su conocimiento primordial para llevar a cabo cualquier estudio en el que muestra y población estén implicados. El *tamaño de la muestra* es, no cabe duda, crucial en todo el proceso, pues a nadie se le escapa que a mayor tamaño, más acertados estaremos en nuestras conclusiones; pero, desgraciadamente, el aumento del tamaño encarece la obtención de la muestra, lo que impide que aquel crezca tanto como desearíamos.

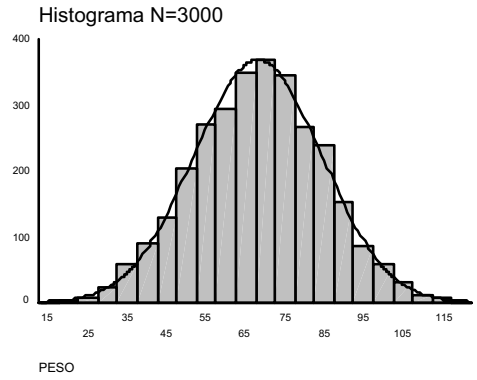
> **La curva de frecuencias** Los histogramas que siguen representan las distribuciones de frecuencias de muestras de *pesos* cuyos tamaños hemos ido aumentando progresivamente. Observamos que a medida que *N* crece los histogramas evolucionan hacia una suavización del contorno superior de sus barras, evolución que nos permite intuir que, para una teórica muestra que contuviera toda la población, el histograma acabaría pareciéndose, si no coincidiendo, a la curva continua que hemos sobrepuesto a las gráficas correspondientes a los tamaños 3000 y 10000. Esta curva límite recibe el nombre de **curva de frecuencias** o **curva de densidad** y tiene

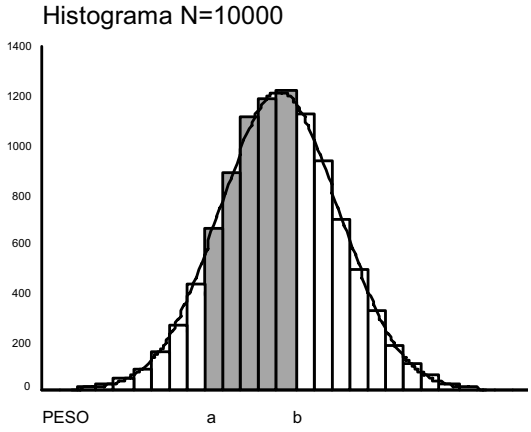
la propiedad de que las frecuencias relativas se representan en ella como área. Así, para cualesquiera dos pesos *a* y *b*, el área que hay bajo la curva entre *a* y *b* es la frecuencia relativa de los pesos que hay entre ambas cantidades.











TEMA 2.- REGRESIÓN Y CORRELACIÓN LINEAL

1. Descripción conjunta de las observaciones de dos variables

El tema 1 desarrollaba métodos gráficos y numéricos para la descripción de datos provenientes de la observación de una variable. Aplicábamos los distintos métodos a las 250 observaciones de una encuesta y aunque las variables observadas eran varias, cada una de ellas era descrita por separado. Aun cuando el análisis conjunto de algunas de estas variables, por ejemplo la altura y el peso, sea razonable y conveniente, no era posible llevarlo a cabo con los métodos entonces descritos.

El objetivo de este tema es proporcionar métodos para analizar la variación conjunta de pares de observaciones pertenecientes a dos variables continuas, con el objetivo de detectar la existencia de algún tipo de dependencia funcional entre ambas. Aunque los posibles tipos de dependencia entre dos variables son muchos, nos ocuparemos solamente del caso lineal, aquel en el que una recta explica suficientemente la relación entre ambas variables. En la primera parte introduciremos características numéricas y métodos de representación gráfica que permitan cuantificar e intuir el grado y tipo de dependencia, dedicando la segunda parte a la obtención de la llamada recta de regresión. Nos valdremos, también ahora, de un ejemplo que facilite la comprensión de los nuevos conceptos.

Altura y peso En la tabla se muestran las alturas (cm.) y los pesos (kg.) de 38 individuos, elegidos al azar, entre los 250 que contestaron la encuesta que introducíamos en le tema 1.

altura	peso	altura	peso
190	80	149	67
155	56	190	93
167	41	162	58
171	49	181	78
182	89	166	69
173	71	160	52
151	53	165	58
172	71	182	86
175	89	151	48
189	93	192	109
162	80	162	39
183	88	162	65
162	65	160	68
173	78	162	63
147	60	200	86
189	85	202	96
185	56	182	84
159	58	150	45
150	55	168	58
Media	$\overline{x}_{altura} = 170.55$		$\overline{x}_{peso} = 69.45$
Desviación típica	$s_{altura} = 15.05$		$s_{peso} = 17.18$

La experiencia demuestra que, en general, las personas altas tienen mayor peso. Veamos cómo poner de manifiesto este hecho a partir de las observaciones anteriores.

Covarianza La covarianza entre dos variables observadas, X e Y, se mediante la expresión

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1},$$

donde n es el número de observaciones. Como en otras ocasiones, existe una expresión alternativa que facilita el cálculo de la covarianza,

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n(n-1)}$$

Para los datos de altura (X) y peso (Y) observados podemos disponer los cálculos de la siguiente forma:

x	y	xy
190	80	15200
155	56	8680
167	41	6847
171	49	8379
182	89	16198
173	71	12283
151	53	8003
172	71	12212
175	89	15575
189	93	17577
162	80	12960
183	88	16104
162	65	10530
173	78	13494
147	60	8820
189	85	16065
185	56	10360
159	58	9222
150	55	8250
149	67	9983
190	93	17670
162	58	9396
181	78	14118
166	69	11454
160	52	8320
165	58	9570
182	86	15652
151	48	7248
192	109	20928
162	39	6318
162	65	10530
160	68	10880
162	63	10206
200	86	17200
202	96	19392
182	84	15288
150	45	6750
168	58	9744
Suma	**6.481**	**2.639**	**457.406**

y, de aquí,

$$s_{xy} = \frac{457406}{37} - \frac{6.481 \times 2.639}{37 \times 38} = 197.77$$

Se supone que este valor nos proporciona información acerca de la relación de dependencia existente entre ambas variables, ¿pero de qué manera lo hace? ¿Cómo interpretar el resultado que acabamos de obtener? Para ello interpretemos la covarianza a través de su **signo** y de su **magnitud**. Como la interpretación requiere de la representación gráfica de las observaciones, hablaremos primero de los llamados **gráficos de dispersión**.

Gráficos de dispersión Una representación gráfica bidimensional de las observaciones permite confirmar visualmente la existencia de una relación de dependencia entre las variables. En algunas situaciones podemos, incluso, intuir la forma de dicha dependencia. Se trata, simplemente, de representar los pares de valores mediante puntos a través de los ejes de coordenadas X e Y, eligiendo adecuadamente las unidades en cada eje, aunque la mayoría de métodos de representación gráfica que existen a nuestra disposición en los ordenadores personales lo hacen de manera automática.
Para los datos de altura y peso, el gráfico de dispersión correspondiente se muestra en la Figura 1, y de él parece deducirse una relación de tipo lineal entre altura y peso.

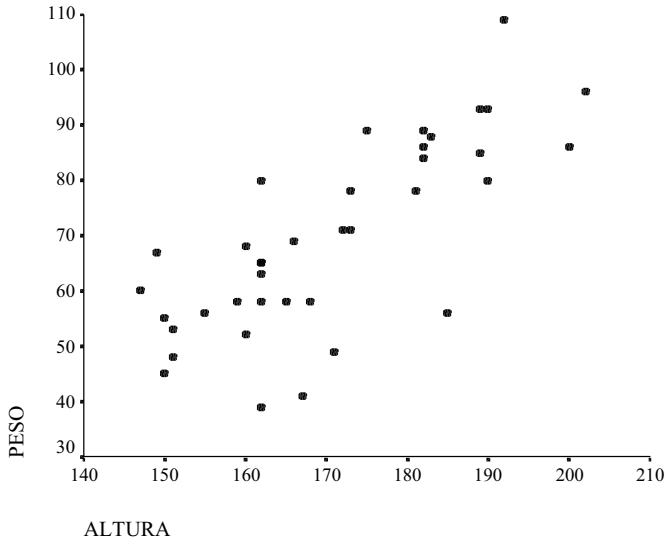


Figura 1.- Gráfico de dispersión correspondiente a las observaciones de altura y peso

Signo de la covarianza A diferencia de lo que ocurría con la varianza, que por tratarse de la media de una suma de cuadrados nunca puede ser negativa, la covarianza puede ser positiva, negativa o nula.

- **Covarianza positiva**: denota una relación **creciente** entre las dos variables, es decir, que cuando una aumenta la otra también lo hace. Este es el caso de la relación existente entre altura y peso, pues es bien sabido que, por regla general, el peso aumenta con la altura.
- **Covarianza negativa**: denota una relación **decreciente** entre las dos variables, es decir, que cuando una aumenta la otra disminuye. El gráfico de dispersión de la Figura 2 nos muestra una relación de este tipo entre la **latitud** y la **temperatura máxima en enero (°F)** en diversas ciudades de EE.UU.

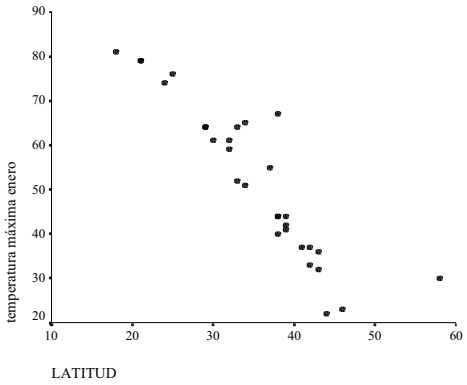


Figura 2.- Gráfico de dispersión correspondiente a las observaciones de latitud y temperatura máxima (°F) en el mes de enero

- **Covarianza nula**: denota, bajo ciertas condiciones, ausencia de cualquier tipo de relación entre ambas variables y, siempre, la ausencia de relación de tipo lineal.

Para justificar las anteriores afirmaciones observemos la gráfica de dispersión correspondiente a las observaciones de alturas y pesos, en la que hemos añadido sendas rectas perpendiculares que se cruzan en el **centro de gravedad** de los datos observados, es decir, el punto de coordenadas $(\overline{x}, \overline{y})$. Estas rectas dividen el plano en cuatro regiones, que aparecen numeradas en la figura.

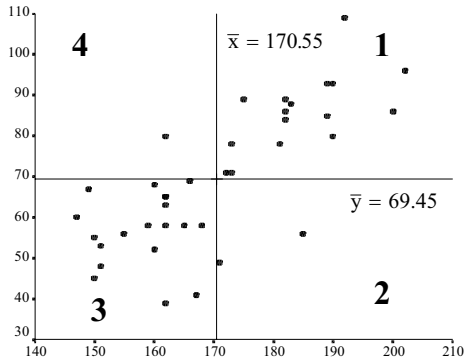


Figura 3.- Cuadrantes de signo para las desviaciones de las variables respecto de sus medias

En cada uno de estos cuadrantes se verifica:

$$\text{- en } \mathbf{1},\ x > \overline{x},\ y > \overline{y} \Rightarrow (x-\overline{x})\cdot(y-\overline{y}) > 0$$

$$\text{- en } \mathbf{2},\ x > \overline{x},\ y < \overline{y} \Rightarrow (x-\overline{x})\cdot(y-\overline{y}) < 0$$

$$\text{- en } \mathbf{3},\ x < \overline{x},\ y < \overline{y} \Rightarrow (x-\overline{x})\cdot(y-\overline{y}) > 0$$

$$\text{- en } \mathbf{4},\ x < \overline{x},\ y > \overline{y} \Rightarrow (x-\overline{x})\cdot(y-\overline{y}) < 0$$

Si la relación que existe entre ambas variables es **creciente**, como es el caso de la gráfica, los puntos de la dispersión estarán mayoritariamente repartidos entre los cuadrantes 1 y 3. Para una relación **decreciente**, esta dispersión se producirá entre los cuadrantes 2 y 4. Cuando los puntos se distribuyan de manera más o menos equilibrada entre los cuatro cuadrantes, la covarianza será muy pequeña porque los productos con signo positivo y negativo tenderán a anularse.

Magnitud de la covarianza En general, podemos afirmar que valores mayores de la covarianza denotan una mayor intensidad de la relación funcional entre las variables. Aunque esta afirmación habrá de ser matizada posteriormente, veamos primero dos ejemplos que la ilustran.
Para la altura y el peso, su gráfico de dispersión (Figura 1) indica la existencia de una relación, probablemente de tipo lineal, que es creciente. Para estos datos el valor de su covarianza era

$$s_{\text{altura,peso}} = 197.77$$

Consideremos ahora los datos de la tabla siguiente, que contiene observaciones correspondientes al precio medio del billete de un autobús urbano (X) y al precio medio del kilo de alcachofas (Y) en 30 capitales de provincia y durante la campaña del invierno 97-98. Ambos precios vienen expresados en pesetas. Puede constatarse que los valores observados, sus medias y sus desviaciones típicas son, todos ellos, del mismo orden de magnitud que los obtenidos para la altura y el peso.

x	y	x	y
170	68	196	81
153	66	193	60
194	41	139	70
170	85	162	67
166	54	173	48
174	76	155	54
166	74	214	79
191	88	143	85
163	68	149	56
149	78	147	85
161	53	157	82
167	23	170	69
166	53	164	64
151	102	157	72
186	52	177	63

	autobús	alcachofas
media	167.43	67.20
desviación típica	17.51	16.21

Si realizamos una representación gráfica de las parejas de valores observados (Figura 4)

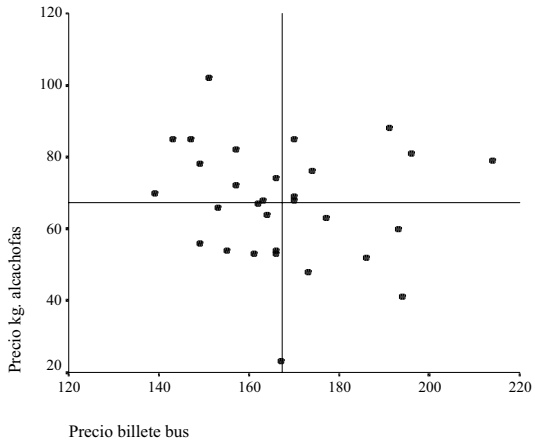


Figura 4.- Gráfico de dispersión correspondiente a los precios del autobús y las alcachofas

constataremos algo que la lógica nos anunciaba, la aparente falta de relación entre ambos tipos de observaciones. El valor de la correspondiente covarianza,

$$s_{bus,alcachofa} = -37.33,$$

casi seis veces menor que la covarianza para altura y peso, confirma lo que visualmente adivinábamos.

Parece pues claro que a mayor valor de la covarianza más fuerte es la relación de dependencia existente entre las variables, pero esta afirmación ha de ser matizada en función de la siguiente propiedad de la covarianza:

> **Propiedad de la covarianza** Si llevamos a cabo una transformación lineal de las variables X e Y,
>
> $$U = aX + b \qquad V = cY + d,$$
>
> la covarianza de las nuevas variables sufre la siguiente transformación:
>
> $$s_{uv} = a \cdot c \cdot s_{xy}$$

Ello supone, por ejemplo, que si expresamos la altura en metros, $U = X/100$, y el peso en arrobas, aunque sea unidad más propia de los gorrinos que de los humanos, $V = Y/12$, tendremos

$$s_{uv} = \frac{1}{12 \cdot 100} \cdot s_{xy} = \frac{197.77}{1200} = 0.16$$

¿Quiere ello decir que por el mero hecho de expresar las variables en otras unidades su relación de dependencia ha cambiado? Como la respuesta es, obviamente, no, esta circunstancia nos lleva a matizar la afirmación que antes hacíamos: *para parejas de observaciones con valores del mismo orden de magnitud, a mayor covarianza, mayor dependencia funcional.*

El matiz, aunque necesario, no nos resuelve la situación que pueda producirse cuando pretendamos comparar las covarianzas de series de datos con valores de muy diferente orden de magnitud. La solución requiere introducir una nueva característica numérica para los pares de valores observados.

Coeficiente de correlación lineal Una forma de evitar el problema anterior, es definir una característica que sea insensible a los cambios de escala. Entre las muchas que podrían introducirse, la más extendida es el llamado **coeficiente de correlación** entre las variables X e Y, r_{xy}. Se define mediante la expresión,

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{s_{xy}}{s_x \cdot s_y}$$

Este coeficiente goza de unas interesantes propiedades que justifican su utilización.

Propiedades del coeficiente de correlación:

PC1) Si $U = aX + b$ y $V = cY + d$, entonces

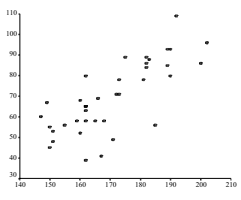
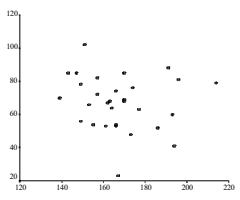
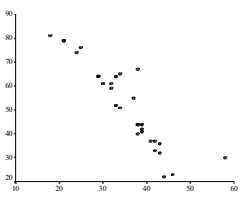
$$r_{uv} = \begin{cases} r_{xy}, & \text{si } a \cdot c > 0 \\ -r_{xy}, & \text{si } a \cdot c < 0 \end{cases}$$

PC2) $-1 \le r_{xy} \le 1$

PC3) Si,

$r_{xy} = 1$, entre X e Y existe dependencia lineal creciente, $Y = aX + b$, con $a>0$,
$r_{xy} = -1$, entre X e Y existe dependencia lineal decreciente, $Y = aX + b$, con $a<0$.

La **primera** de estas propiedades resuelve el problema que se nos había planteado con el cambio de valor que los cambios de escala producen en la covarianza. A lo sumo cambiará el signo del coeficiente, dependiendo esto a su vez de los signos que tengan los cambios de escala introducidos, a y c.
Las propiedades **segunda** y **tercera**, nos dicen que $|r_{xy}|$ describe el grado de linealidad existente entre X e Y, en una escala que va de 0 a 1, indicando el valor 0 la ausencia de relación lineal y el valor 1 la existencia de una relación lineal perfecta. Si los valores de r_{xy} son negativos, indican dependencia decreciente, una variable crece mientras la otra decrece o viceversa, mientras que valores positivos de r_{xy} indican que esta relación es creciente.
Los valores de los coeficientes de correlación de los datos correspondientes a los tres ejemplos anteriores y sus gráficos de dispersión nos ayudarán a ilustrar y comprender estas propiedades.

altura-peso	autobús-alcachofa	latitud-temperatura
$r_{xy} = 0.76$	$r_{xy} = -0.13$	$r_{xy} = -0.90$
		

2. Recta de regresión de Y sobre X

Hemos hablado en el apartado anterior de relación funcional entre las variables X e Y y hemos dicho que ésta puede de ser de muy diversos tipos. En este apartado nos vamos a ocupar de estudiar aquella situación en la que una recta describe adecuadamente la dependencia entre ambas.
Antes de describir la obtención de la recta más conveniente a nuestros datos, conviene que comencemos explicando cuál es el significado de la recta de regresión y el objetivo que se persigue con su obtención. Asumida la existencia de una relación lineal entre las variables que hemos observado, el **ajuste**, así se denomina el proceso, de una recta de regresión a nuestros datos pretende dotarnos de un modelo teórico que describa, lo mejor posible, la dependencia observada. El objetivo que perseguimos al disponer de una recta que se ajusta bien a nuestros datos, es poder llevar a cabo **predicciones** de la variable Y a partir de valores predeterminados de la variable X. Por ejemplo, entre las observaciones de alturas y pesos no existen ninguna que corresponda a una altura de 178 cm., la recta de regresión ajustada puede predecir qué peso correspondería a esta altura sin más que sustituir el valor x =178 en la ecuación de la recta.
Recordemos que la forma más sencilla de la ecuación de una recta es

$$Y = a\,X + b$$

y, en consecuencia, nuestro objetivo será encontrar los valores de los parámetros de la recta, a y b, que reciben el nombre de **pendiente** y **ordenada en el origen**, respectivamente. Estos valores dependerán del **criterio** con el que la recta se elija y el problema estriba en que son muchos los posibles criterios a utilizar. Por ejemplo:

- **C1** **Puntos extremos** La recta ajustada con este criterio pasaría por el punto más bajo (menor valor de y) y más a la izquierda (menor valor de x) y por el más alto (mayor valor de y) y más a al derecha (mayor valor de x).
- **C2** **Igual reparto** La recta ajustada con este criterio pasaría por el centro de gravedad de los datos observados, $(\bar{x}, \bar{y})$, y dejaría a cada lado la mitad de las observaciones.
- **C3** **Mínimas distancias** La recta se elige de tal forma que la suma de los cuadrados de las distancias de cada punto a la recta es mínima.
- **C4** **Mínimos cuadrados** En las observaciones tenemos parejas de valores (x_i,y_i). La recta obtenida bajo este criterio, minimiza la suma de los cuadrados de las diferencias entre el valor de y_i observado y el obtenido al sustituir en la ecuación de la recta el valor x por x_i.

No todos estos criterios actúan con la misma bondad, basta observar el resultado a que algunos de ellos de conducen en la gráfica que sigue. En ella hemos sobrepuesto al gráfico de dispersión de los datos altura-peso las rectas correspondientes a C1 y C2.

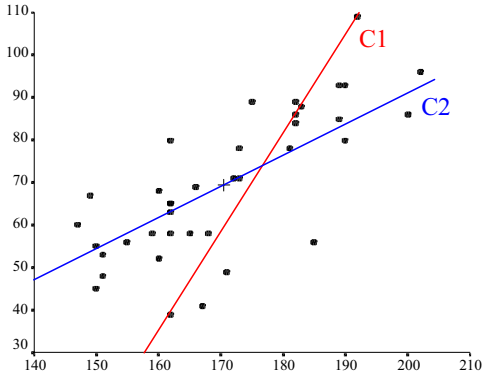


Figura 5.- Rectas correspondientes a los criterios C1 y C2 para los datos de altura y peso

El criterio C1 no parece producir un buen ajuste, mientras que la recta correspondiente a C2 goza de mejor calidad., pero tiene el inconveniente de ser un método gráfico poco eficiente e impreciso porque la recta a determinar no es única.

El criterio de los mínimos cuadrados es el habitualmente utilizado porque da lugar a una recta con buenas propiedades, permite obtener sencillas expresiones para los parámetros de la recta y guarda una estrecha e interesante relación con el coeficiente de correlación. Vamos pues a ocuparnos de él con más detalle.

Recta de regresión mínimo-cuadrática Recordaremos nuevamente en qué consiste el ajuste de una recta mediante el método de los mínimos cuadrados apoyándonos en una gráfica que nos facilite su comprensión.

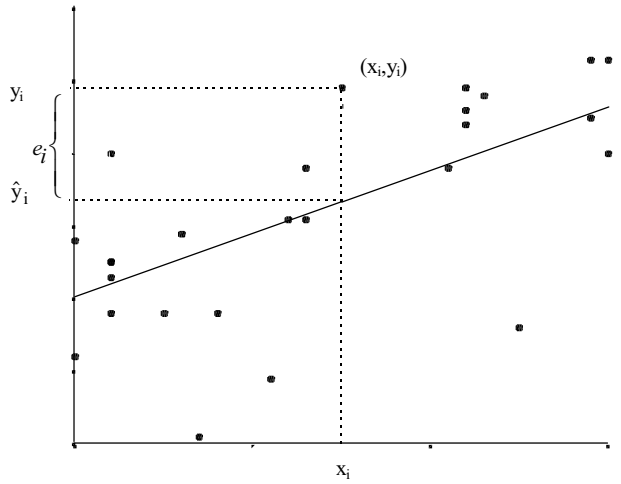


Figura 6.- Residuo en una recta de regresión

En la Figura 6 hemos representado un gráfico de dispersión cualquiera. En él observamos, que a la pareja de datos (x_i, y_i) podemos hacerle corresponder sobre la recta otro punto cuyas coordenadas son $(x_i, \hat{y}_i)$, siendo $\hat{y}_i$ la predicción que la recta nos da para x_i. Entre esta predicción y el valor observado existe una diferencia que denominamos **residuo** o **error**,

$$e_i = y_i - \hat{y}_i$$

El **método de los mínimos cuadrados** consiste en encontrar valores para los parámetros de la recta, *a* y *b,* tales que la llamada *suma de cuadrados de los errores*,

$$SC_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (ax_i - b))^2 \ ,$$

sea mínima.

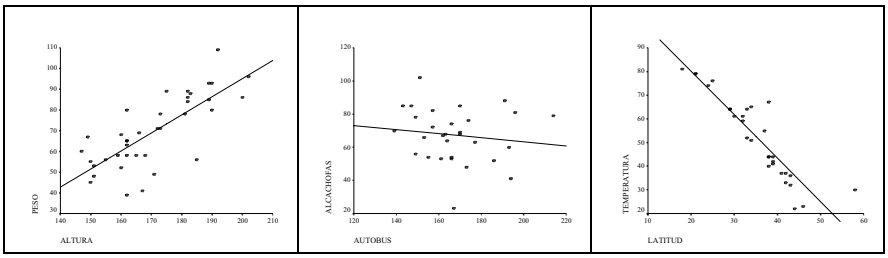
Con esta condición, los valores para *a* y *b* vienen dados por las expresiones:

$$a = \frac{s_{xy}}{s_x^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}, \qquad b = \bar{y} - a\bar{x}$$

La recta de regresión así obtenida pasa por $(\bar{x}, \bar{y})$, centro de gravedad de los datos observados, como se deduce del valor de la ordenada en el origen, *b*.

Obtengamos ahora las rectas de regresión para los tres conjuntos de datos que venimos manejando.

altura-peso	p_autobús-p_alcachofa	latitud-temperatura
y = 0.87x - 79.32	y = -0.12x + 85.59	y = -1.83x + 116.75



Como ya sabíamos, el ajuste es tanto mejor cuanto mayor es el valor absoluto del coeficiente de correlación. En el caso de los precios del autobús y de las alcachofas, la recta no parece ser un buen modelo para describir la dependencia entre ambos, si es que existe. Pero no podemos juzgar la bondad del ajuste de manera empírica solo mediante la observación de las gráficas. ¿Es posible *medir* la calidad del ajuste? Para responder a esta pregunta estudiaremos el cociente entre la varianza de los valores observados de *Y* y la varianza de los errores o residuos y relacionaremos dicho cociente con el coeficiente de correlación.

Cociente entre s_y^2 y s_e^2 Comencemos puntualizando que por s_e^2 designamos la varianza de los errores, que se obtiene a partir de la expresión,

$$s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1},$$

pero, como fácilmente puede comprobarse, $\bar{e} = 0$, lo que reduce la expresión a

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-1} = \frac{SC_e}{n-1}$$

¿Qué interés tiene para nosotros el cociente entre ambas varianzas? Recordemos que el objetivo perseguido con la obtención de la recta de regresión mínimo-cuadrática es, en la medida que se ajusta bien a las observaciones, dotarnos de un modelo que nos permita **predecir** el valor de *y* asociado a un valor cualquiera *x*. Es posible efectuar dicha predicción a partir de los propios datos observados sin necesidad de ajustar recta alguna. En efecto, puesto que la media de un conjunto de observaciones tiene carácter representativo de las mismas, podemos tomarla como predicción para cualquier valor de *x*.

Si actuamos así, ¿qué error total estamos cometiendo? Una medida de ese error, a semejanza de lo que hemos hecho con los errores o residuos obtenidos a partir de la recta de regresión, puede obtenerse utilizando el cuadrado de la diferencia entre el valor observado, y_i, y la predicción, $\bar{y}$, lo que nos conduce a la varianza de las observaciones:

$$error\ total\ con\ \bar{y} = s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Cuando esta misma predicción la llevamos a cabo con la recta de regresión obtenida, el error total cometido será la varianza de los residuos, s_e^2, cuya expresión acabamos de dar.

La obtención de la recta de regresión tiene validez en la medida que reduzca el error. Lo mejor será conocer la proporción de reducción que hemos llevado a cabo al utilizar la recta para predecir. Una manera sencilla de hacerlo es utilizar la expresión,

$$reducción\ del\ error = \frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Podemos volver sobre nuestros tres ejemplos y calcular esta reducción cuando utilizamos las rectas de regresión que hemos ajustado a cada caso. La tabla recoge los cálculos y muestra el % de reducción en la última columna. Como era de prever, la mayor reducción se obtiene para las observaciones de latitud y temperatura, con un 82%, para la altura y el peso dicha reducción es casi del 60%, mientras que para el precio del autobús y el de las alcachofas es prácticamente inexistente, menos del 2%.

Reducción de la varianza

	s_y^2	s_e^2	reducción	%
altura-peso	287.46	119.49	0.5843	58.43%
autobús-alcachofas	253.89	249.50	0.0173	1.73%
latitud-temperatura	288.49	52.15	0.8192	81.92%

Es posible representar gráficamente el efecto que la recta tiene en la reducción. Para ello representamos conjuntamente las diferencias entre observaciones y predicciones en ambos casos tal y como hemos hecho en la Figura 7. En todas las gráficas la parte superior representa mediante un trazo, para cada valor de *x*, la diferencia entre la *y* observada y la media, que ha sido representada mediante una recta. Los trazos por debajo de media indican que la diferencia es negativa. La parte inferior representa las diferencias (errores o residuos) entre el valor observado de *y* y el obtenido a partir de la recta mediante la sustitución del correspondiente *x*. También ahora hemos dibujado la recta correspondiente a la media de estos errores que, recordemos, vale 0. Una vez más, la gráfica es elocuente en los dos casos extremos: latitud-temperatura y autobús-alcachofas.

Figura 7.-Gráfica de las diferencias entre la predicción y el valor observado del peso cuando aquella se lleva a cabo con la media (superior) o con la recta de regresión (inferior)

Figura 8.-Gráfica de las diferencias entre la predicción y el valor observado del precio del kilo de alcachofas cuando aquella se lleva a cabo con la media (superior) o con la recta de regresión (inferior)

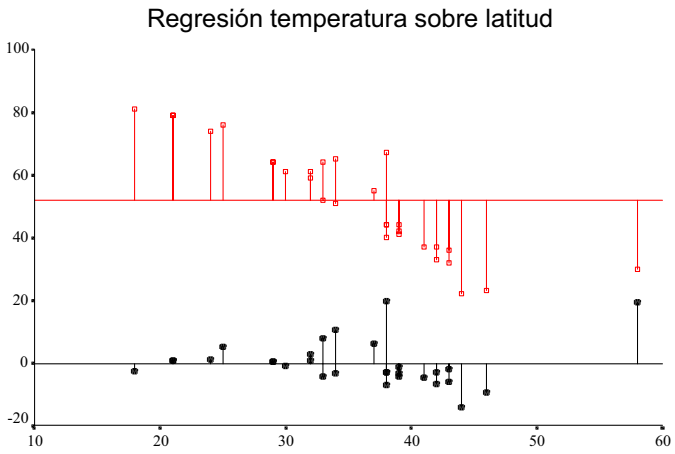


Figura9.-Gráfica de las diferencias entre la predicción y el valor observado de la temperatura cuando aquella se lleva a cabo con la media (superior) o con la recta de regresión (inferior)

Regresión y correlación Ya hemos dicho que el coeficiente de correlación mide, en una escala de 0 a 1, el grado de linealidad existente entre ambas variables. Pero no solo eso, sino que además nos proprociona información acerca de la reducción de varianza conseguida mediante la recta de regresión. En efecto, en la tabla de reducción de la varianza anteriormente obtenida, vamos a incluir el valor del coeficiente de correlación y de su cuadrado.

	r	**reducción**	**r^2**
altura-peso	0.7644	0.5843	0.5843
autobús-alcachofas	-0.1316	0.0173	0.0173
latitud-temperatura	-0.9051	0.8192	0.8192

Comprobamos que dicho cuadrado coincide, en todos los casos, con la reducción de varianza obtenida. Este resultado no es casual y responde a una conocida propiedad que relaciona correlación y regresión a través de la siguiente expresión,

$$r_{xy}^2 = 1 - \frac{s_e^2}{s_y^2}$$

Este resultado hace innecesario cualquier cálculo adicional para conocer la reducción de varianza que el ajuste de una recta de regresión comporta. Basta con obtener el cuadrado del coeficiente de correlación, r_{xy}^2, que es conocido como el **coeficiente de determinación**.

3. Un comentario final

La presentación que hemos hecho en este tema es puramente descriptiva. Pero, ¿qué ocurre cuando los datos provienen de una muestra de variables aletorias? En ese supuesto, todas las características implicadas son también aleatorias y en particular dos de ellas merecen especial interés: el coeficiente de correlación y la recta de regresión a través de los parámetros que la definen, a y b. Es posible en este contexto llevar a cabo contrastes de hipótesis acreca de todas estas variables aleatorias. En el caso de r_{xy} el contraste más habitual consiste en H_0: $r_{xy} = 0$, frente a H_A: $r_{xy} \neq 0$, mientras que para la recta de regresión se plantea también contrastar si sus parámetros son nulos. El desarrollo de estos contrastes queda fuera del alcance y objetivos de este curso.

TEMA 3.- PROBABILIDAD

1. Definición y Propiedades

A la pregunta *¿cuánto tiempo tardará un automóvil en recorrer los 350 km que separan Valencia de Madrid si hace el recorrido a una velocidad de 100 km/hora?*, respondemos sin dudar que tardará 3 horas y media. La situación es más complicada cuando alguien nos pregunta qué cara mostrará la moneda que va a lanzar. En este segundo caso podemos responder que será *cara* o *cruz* y podemos añadir, porque la experiencia nos lo indica, que ambos resultados tienen la misma "probabilidad" de aparecer, porque suponemos que se trata de una moneda correcta como suele serlo prácticamente siempre.

¿Qué diferencia hay entre ambos fenómenos? En el primer caso estamos ante un fenómeno **determinista**, en los que la relación entre la causa y el efecto está determinada, es conocida a través de una ecuación, $e = v \cdot t$, que liga el espacio, la velocidad y el tiempo. La llamadas Ciencias Experimentales se ocupan de estudiar este tipo de fenómenos y de establecer los *modelos deterministas* (ecuaciones) que rigen su comportamiento. En el segundo caso el fenómeno es **aleatorio**, al igual que el lanzamiento de un dado o la extracción al azar de una bola de una urna, y de él sólo sabemos el conjunto de posibles resultados que pueden darse. Si queremos conocer más acerca de este tipo de fenómenos hemos de recurrir a la **Teoría de la Probabilidad**, que los estudia para poder establecer *modelos probabilísticos* que nos permitan predecir sus resultados y poder afirmar, por ejemplo, que al lanzar un dado nos puede aparecer cualquiera de los números 1, 2, 3, 4, 5 o 6, con igual probabilidad, 1/6.

En todo fenómeno aleatorio hay que distinguir cuatro elementos esenciales: **experimento aleatorio**, que es la prueba que llevamos a cabo, **resultado**, **espacio muestral**, que es el conjunto de todos los posibles resultados a los que nos conduce el experimento y **suceso**, que es cualquier subconjunto de resultados caracterizados por determinada propiedad. La Tabla 1 recoge algunos ejemplos sencillos en los que se describen cada uno de estos elementos.

	Experimento	Espacio muestral	Algunos sucesos
E1	Lanzar dos monedas	S={CC,C+,+C,++}	A={Ha salido una cara}={C+,+C} B={Ha salido más de una cruz}={++}
E2	Lanzar un dado	S={1,2,3,4,5,6}	A={La cara es par}={2,4,6} B={La cara es mayor que 5}={6}
E3	Llamadas a una centralita telefónica	S={0,1,2,3,4,...}	A={Llega alguna llamada}={1,2,3,...} B={No llegan más de 7 llamadas} ={0,1,2,3,4,5,6,7}
E4	Sexo de los 3 hijos de un matrimonio	S={VVV,VVM,VMV,MVV, VMM,MVM,MMV,MMM}	A={El matrimonio solo tiene varones} ={VVV} B={El matrimonio tiene al menos una niña} = ={VVM,VMV,MVV,MMV,MVM,VMM,MMM}
E5	Elegir al azar un punto en un círculo de radio 1	S={Los puntos del círculo}	A={El punto dista del centro menos de 0.5} B={La cuerda que pasa por el punto es menor que 1}
E6	Elegir al azar una persona con 15 o más años	S={Las personas de 15 o más años}	A={La persona es mujer} B={La persona está a favor de la semana de 35 horas}

Tabla 1.- **Algunos ejemplos de experimentos aleatorios**

Como podemos comprobar en la tabla anterior, los experimentos aleatorios pueden dar lugar a espacios muestrales diversos, desde espacios con un número finito de puntos o resultados, hasta espacios con una cantidad no numerable de puntos (caso del círculo). Es costumbre designar el espacio muestral mediante ***S***, el resultado con ***s*** y los sucesos mediante las mayúsculas de la primeras letras del abecedario, ***A***, ***B***, ...

Cuando el resultado de nuestro experimento es un punto que está en algún suceso, decimos que dicho suceso se **ha realizado** o **ha ocurrido**. Así, en E2, si al lanzar el dado nos aparece la cara 6, diremos que ha ocurrido A porque 6 es par, pero también ha ocurrido B porque 6 es mayor que cinco.

Algunos sucesos con nombre propio En cualquier experimento existen siempre unos sucesos que por su importancia tienen nombre propio, son los llamados **suceso cierto** y **suceso imposible**. El primero es el propio espacio muestral, S, y se llama cierto porque siempre ocurre puesto que el resultado de nuestro experimento es siempre un punto de S, ya que éste los contiene a todos. El suceso imposible es, por el contrario, aquel que no ocurre nunca, cosa que solo puede suceder si no contiene ningún punto, razón por la cual se le designa con el símbolo del conjunto vacío, $\varnothing$. Obsérvese que cualquier otro suceso estará comprendido entre ambos.
En cualquier experimento, al definir un suceso A estamos simultáneamente definiendo otro suceso: el suceso constituido por todos los puntos que no están en A, que llamaremos **complementario** de A y designaremos mediante A^c. Este suceso representa la negación de A y ocurre siempre que no lo hace A. Por ejemplo, en E2, si definimos A=*{la cara es par}*, A^c=*{la cara es impar}*. Obsérvese que el resultado del experimento pertenece siempre a A o A^c, lo que implica que siempre se realiza uno de los dos.

Sucesos y conjuntos Hemos visto que los sucesos no son más que subconjuntos de S y por tanto podemos efectuar con ellos las operaciones que llevamos a cabo con los conjuntos: **unión** e **intersección**. La unión de dos sucesos, $A \cup B$, es un nuevo suceso que contiene los resultados de ambos y que se realiza cuando lo hacen A o B o ambos a la vez. Por ejemplo, en E2 si definimos *C={la cara es menor o igual que 3}={1,2,3}*, el suceso $A \cup C$ =*{1,2,3,4,6}*. La intersección de dos sucesos, $A \cap B$, es un nuevo suceso que contiene los resultados que son comunes a ambos. En el caso anterior, $A \cap C$ =*{2}*, y en E6, $A \cap B$ =*{las mujeres que están a favor de la semana de 35 horas}*. Observemos que cuando dos sucesos comparten puntos pueden realizarse simultáneamente, como ya vimos antes con los sucesos A y B de E2. Decimos entonces que son **compatibles**. Pero en ocasiones los sucesos no tienen puntos comunes, sería por ejemplo el caso de los sucesos A y B de E4, su intersección no contiene ningún punto y es por tanto el suceso imposible, $\varnothing$, que no puede realizarse. Decimos en este caso que los sucesos son **incompatibles**, lo que supone que no pueden ocurrir simultáneamente. Un claro ejemplo de sucesos incompatibles son A y A^c.

Muestreo aleatorio En algunos de los ejemplos que recoge la Tabla 1, hemos descrito nuestro experimento mediante la expresión *"Elegir al azar ..."*, que nos resulta familiar por ser muy utilizada en el lenguaje habitual y porque los juegos de azar, el referente más inmediato, forman parte de nuestra cultura cotidiana. Todos hemos visto alguna vez la retransmisión de un sorteo de la Lotería Nacional o de la Lotería Primitiva en los que se llevan a cabo extracciones al azar de bolas numeradas que configuran los números que obtendrán premio. Somos conscientes de que el éxito de estos juegos de azar radica en la confianza, ratificada por la experiencia de muchos años, que todos aquellos que juegan tienen en la *corrección* de los sorteos. La corrección supone que las cosas transcurren de tal modo que ninguna de las bolas predomina sobre las otras a la hora de ser extraída y que la extracción de una bola no influye para nada en cuáles serán las restantes. En este mismo principio se basa lo que conocemos como **muestreo aleatorio**, proceso que consiste en obtener una muestra de tamaño *n* de una población de acuerdo a las siguientes reglas:

1. todos los miembros de la población tienen la misma posibilidad de ser elegidos para formar parte de la muestra, y
2. los elementos de la muestra son elegidos independientemente unos de otros, es decir, que la presencia de cualquiera de ellos en la muestra no influye en la futura presencia de los demás.

Cuando en adelante utilicemos la expresión *elegir al azar*, estamos haciendo referencia a situaciones en las que se cumplen las anteriores condiciones.

En párrafos anteriores hemos introducido la infraestructura necesaria para definir el concepto de probabilidad: espacio muestral y sucesos. Lo que perseguimos es poder responder a la pregunta que nos planteábamos al principio y que ahora repetimos de forma más general: *si llevamos a cabo un experimento aleatorio y A es un suceso ligado a él, ¿cuál es la probabilidad de A?* Pero hay preguntas que generan más preguntas en lugar de respuestas. La pregunta anterior hace surgir de inmediato otra, *¿cómo obtener dicha probabilidad?* A continuación describimos dos de los procedimientos más conocidos para asignar probabilidades.

Método frecuencialista: la probabilidad como límite de la frecuencia Bajo ciertas condiciones es posible identificar probabilidad y frecuencia relativa. Se trata de aquellas situaciones en las que nuestro experimento es susceptible de ser repetido infinidad de veces en las mismas condiciones y nos es posible observar en cada repetición la realización o no del suceso que nos interesa, entonces:

la probabilidad de A, P(A), se interpreta como la cantidad a la que se aproxima la frecuencia relativa de ocurrencias de A cuando repetimos indefinidamente el experimento.

Tratemos de comprender mejor esta interpretación mediante el siguiente ejemplo:

Ejemplo Tenemos una bolsa con 100 bolas, **30 negras** y **70 blancas** y extraemos al azar (recordemos lo que ello significa) una de ellas. Si *A={la bola es negra},* queremos conocer la probabilidad de que A ocurra. Para ello procedemos como sigue:

1. llevamos a cabo una extracción, comprobamos el color de la bola y metemos la bola nuevamente en la bolsa,
2. repetimos el proceso *n* veces y, a medida que *n* varia, estudiamos como evoluciona el cociente *m/n*, siendo *m* el número de veces que ha ocurrido A lo largo de las *n* extracciones (obsérvese que el cociente no es más que la frecuencia relativa de ocurrencias de A),
3. representamos gráficamente los valores *n* (abcisas) y *m/n* (ordenadas).

En la Tabla 2 se recogen, parcialmente, los resultados de las primeras 1000 extracciones y la representación gráfica de los valores de *m/n*.

n	color	m	m/n
1	negra	1	1,000
2	negra	2	1,000
3	blanca	2	0,667
4	blanca	2	0,500
5	blanca	2	0,400
……	……	……	…
10	blanca	3	0,300
……	……	……	…
100	blanca	32	0,320
……	……	……	…
500	blanca	149	0,298
……	……	……	…
1000	negra	292	0,292

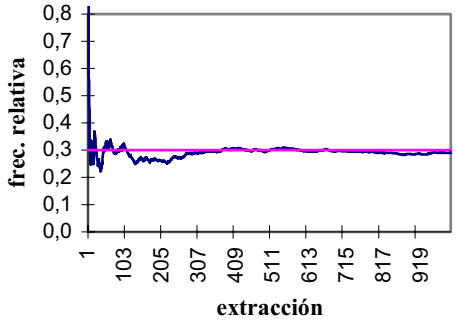


Tabla 2.- **Resultados de las 1000 extracciones de bolas de una bolsa 30N, 70B**

Como puede observarse en la gráfica, la frecuencia relativa de ocurrencias de A oscila alrededor de 0,3 a medida que *n* aumenta y estableceremos que P(A)=0,3. Pero 0,3 es, por otra parte la frecuencia relativa o proporción de bolas negras que existen en la bolsa.

Esta última coincidencia no es tal, pues podemos comprobar empíricamente que cuando la elección está hecha al azar, es decir, según las reglas del muestreo aleatorio*, la probabilidad de que el elemento elegido tenga determinada característica es igual a la frecuencia relativa (proporción) de dicha característica en la población*.

Método clásico: fórmula de Laplace Hay otra situación en la que es sencillo obtener la probabilidad de un suceso. Se trata de aquellos experimentos que conducen a un espacio muestral con un número finito de resultados, $S=\{s_1, s_2, \ldots, s_n\}$, todos ellos igualmente probables. En este contexto, si A es un suceso que contiene *m* de los *n* resultados, P(A) se obtiene a partir de la fórmula,

$$P(A) = \frac{m}{n},$$

propuesta por Laplace a finales del siglo XVIII y que se interpreta diciendo que

La probabilidad de un suceso A es el cociente entre los resultados o ***casos favorables*** *a su realización y el total de resultados o* ***casos posibles****.*

Ejemplo 1 En el experimento E1 de la Tabla 1, consistente en lanzar dos monedas, el espacio muestral estaba constituido por S={CC,C+,+C,++} y si las monedas son correctas, cosa que hemos de suponer mientras no se nos diga lo contrario, todos ellos son igualmente probables, por lo que será de aplicación la fórmula de Laplace. Así, si *A={Ha salido una cara}={C+,+C}*,

$$P(A) = \frac{2}{4} = \frac{1}{2}.$$

Ejemplo 2 Supongamos ahora un experimento consistente en lanzar dos dados simultáneamente. El espacio muestral estará constituido por las 36 parejas de valores resultantes de combinar cada una de la 6 caras de un dado con las 6 caras del otro, a saber,

S={(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),
(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),
(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}.

Como suponemos que los dados no están cargados, nada se ha dicho en contra, todos los resultados son igualmente probables y nuevamente podemos utilizar la fórmula de Laplace. Sea *A={la suma de las caras es menor o igual que 5}* entonces A contiene los puntos, *A={(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)}* y P(A)=10/36=5/18.

Los dos métodos de asignar probabilidades que hemos expuesto no cubren todos los posibles fenómenos aleatorios que puedan surgirnos, pero serán suficiente para cubrir nuestro objetivo de introducir el concepto y familiarizarnos con él.

Independientemente del método elegido y del experimento y suceso que estemos estudiando, la esencia del proceso de obtener la probabilidad de un suceso consiste en asignarle una cantidad entre 0 y 1. Pero esta cantidad no puede asignarse de cualquier manera, existen reglas o condiciones que las probabilidades deben de verificar para que alcancen el objetivo de proporcionarnos modelos que nos describan y expliquen los fenómenos aleatorios. Se impone pues dar una definición formal del concepto:

Definición de probabilidad La probabilidad, P, es una función que a cada suceso de un espacio muestral S le asigna una número real, verificándose los siguientes axiomas:

1. Para cualquier suceso A, $P(A) \geq 0$,
2. Para el suceso cierto S, $P(S) = 1$,
3. Si A y B son dos sucesos incompatibles, $A \cap B = \varnothing$, entonces
$$P(A \cup B) = P(A) + P(B).$$

Comentario a la definición No es casualidad que la probabilidad deba cumplir con los axiomas anteriores. Si el método frecuencialista asimila probabilidad a frecuencia relativa, la definición de aquella debe hacerse de manera que exista coherencia entre sus propiedades y las de la frecuencia relativa y, como fácilmente puede comprobarse, los axiomas no son más que una simple transposición de las propiedades que verifica la frecuencia.

Propiedades Para sucesos cualesquiera, A y B, se verifica:

1. $P(\varnothing) = 0$
2. La probabilidad es una función **monótona**, es decir, si $A \subseteq B$ entonces $P(A) \leq P(B)$
3. $0 \leq P(A) \leq 1$
4. $P(A^c) = 1 - P(A)$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Veamos algunos ejemplos en los que aplicar estas propiedades.

Ejemplo 3 Retomemos de nuevo el espacio muestral del ejemplo 2, resultante de observar los resultados de lanzar dos dados. Consideremos los sucesos

A={el producto de ambas caras es 12}={(2,6),(6,2),(3,4),(4,3)}

y B={la suma de ambas caras es 7}={(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)}
y obtengamos P(A), P(B), P(A∪B), y P(B^c).

Como es aplicable la fórmula de Laplace,

$$P(A) = \frac{4}{36} = \frac{1}{9}, \qquad P(B) = \frac{6}{36} = \frac{1}{6}, \qquad P(A^C) = 1 - P(A) = \frac{8}{9}.$$

Para obtener P(A∪B) necesitamos conocer A∩B={(3,4),(4,3)} y por tanto,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{36} + \frac{6}{36} - \frac{2}{36} = \frac{8}{36} = \frac{2}{9}.$$

Ejemplo 4 Las personas de la muestra de tamaño 250 cuyos datos figuran en el Anexo I del Tema 1, se distribuyen con arreglo a su sexo y su estado civil tal como muestra la siguiente tabla:

		ESTADO CIVIL				
		casado/a	**soltero/a**	**viudo/a**	**sep/div**	**Total**
SEXO	**mujer**	63	50	10	3	126
	hombre	62	51	9	2	124
	Total	**125**	**101**	**19**	**5**	**250**

Llevamos a cabo la extracción al azar de una de estas personas y consideremos los siguientes sucesos:

M={la persona es una mujer} C={la persona está casada}
H={la persona es un hombre} S={la persona está soltera}
V={la persona está viuda} D={la persona está separada o divorciada}.

Queremos conocer las probabilidades de los sucesos: M∩C={mujer casada}, H∩V={hombre viudo}, S^c={la persona no está soltera} y H.

Siendo la extracción al azar, la probabilidad de que nuestra persona tenga determinada característica es igual a la frecuencia relativa o proporción de dicha característica. Así, de acuerdo con la tabla,

$$P(M \cap C) = \frac{63}{250} = 0{,}252 \quad P(H \cap V) = \frac{9}{250} = 0{,}036 \quad P(H) = \frac{124}{250} = 0{,}496$$

Para obtener $P(S^c)$ recurriremos a su relación con P(S) y como el número total de solteros es 101,

$$P(S) = \frac{101}{250} = 0{,}404$$

y $P(S^c) = 1 - P(S) = 1 - 0{,}404 = 0{,}596$.

2. Probabilidad condicional

Volvamos a la tabla del ejemplo 4 e imaginemos que antes de obtener P(H) se nos proporciona una ayuda. Se nos dice que la persona extraída al azar está separada o divorciada. Parece lógico que tengamos en cuenta esta información a la hora de obtener P(H), ¿cómo?. Podemos razonar ahora de la siguiente forma: puesto que nos dicen que se trata de una persona separada o divorciada, sólo puede ser una de las 5 que cumplen esta condición, y de éstas 2 son hombres, así pues 2/5 es la proporción correspondiente y la probabilidad que me han pedido es 2/5=0,4.

En esta respuesta lo primero que llama la atención es que la nueva probabilidad es distinta de la anterior que, recordemos, valía 0,496; pero es lógico que así sea porque están obtenidas en situaciones diferentes; en la primera no poseíamos ninguna información *a priori* y nuestro espacio muestral eran las 250 personas sobre las que realizábamos la extracción; en la segunda, puesto que se nos proporciona la información de que el suceso D={la persona es separada o divorciada} ha ocurrido, nuestro espacio muestral se restringe al colectivo de las 5 personas separadas o divorciadas.

Puesto que se trata de probabilidades obtenidas para el mismo conjunto, pero en circunstancias distintas que las hacen a las dos igualmente válidas, deberíamos expresarlas de forma distinta. A esta

nueva probabilidad, obtenida sabiendo que se ha realizado D, la designaremos mediante **P(H|D)** y diremos que es *la probabilidad del suceso H condicionada al suceso D, o sabiendo que se ha realizado el suceso D*.

Volvamos sobre el valor obtenido para P(H|D) y hagamos algunas operaciones sencillas,

$$P(H|D) = \frac{2}{5} = \frac{2/250}{5/250} = \frac{P(H \cap D)}{P(D)},$$

que nos permiten deducir una igualdad que vamos a utilizar para definir el concepto de probabilidad condicionada.

Definición de probabilidad condicionada Dados dos sucesos cualesquiera, A y B, definimos la *probabilidad de A condicionada a B*, P(A|B), mediante:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

donde P(B)>0.

Veamos algunos ejemplos.

Ejemplo 5 Retomando el ejemplo 4, vamos a obtener P(S|H), probabilidad de que la persona extraída sea soltera si sabemos que es un hombre. Aplicando la definición,

$$P(S|H) = \frac{P(S \cap H)}{P(H)} = \frac{51/250}{124/250} = \frac{51}{124} = 0{,}411$$

análogamente, P(H|C), probabilidad de hombre condicionada a que esté casado, es

$$P(H|C) = \frac{62/250}{125/250} = \frac{62}{125} = 0{,}496\,.$$

Ejemplo 6 Supongamos que tenemos una urna con 100 bolas, 80 de las cuales son blancas y las restantes 20 rojas. Nuestro experimento consiste en extraer al azar dos bolas consecutivamente y sin reemplazamiento, es decir, que cada bola extraída no se reintegra a la urna. ¿Cuáles son las probabilidades de que :

a) la primera bola extraída sea roja?
b) la segunda bola sea roja sabiendo que la primera también lo ha sido?

Designemos por R_1 y R_2, respectivamente, los sucesos "las bolas extraídas en primer y segundo lugar son rojas".

a) Se nos pide que calculemos $P(R_1)$. Como la extracción es al azar, la probabilidad es igual a la proporción de bolas del color deseado, así pues,

$$P(R_1) = \frac{20}{100} = 0{,}2$$

b) Se nos pide ahora obtener $P(R_2|R_1)$, para ello razonemos como sigue: al extraer una primera bola roja y no haberla reemplazado, la composición de la urna ha cambiado doblemente, en primer lugar hay ahora 99 bolas porque la primera no ha sido reemplazada, en segundo lugar el número de bolas rojas es 19 porque la primera fue roja. En consecuencia,

$$P(R_2|R_1) = \frac{19}{99} = 0{,}192$$

Supongamos que nos hubieran pedido, en el ejemplo anterior, que obtuviésemos $P(R_2)$. ¿Cómo lo habríamos hecho? Veámoslo:

La bola extraída en primer lugar puede ser roja, suceso R_1 , o blanca, suceso B_1 , cubriendo ambos sucesos todas las posibilidades existentes. Ello supone que $R_1 \cup B_1 = S$ y, como además la bola puede ser blanca o roja, pero no ambas cosas a la vez, se verifica también que son incompatibles, $R_1 \cap B_1 = \varnothing$. Por otra parte, como cualquier suceso está incluido en el espacio muestral, tenemos que $R_2 \cap S = R_2$. Estos resultados, y unas sencillas operaciones con conjuntos, nos permiten escribir:

$$R_2 = R_2 \cap S = R_2 \cap (R_1 \cup B_1) = (R_2 \cap R_1) \cup (R_2 \cap B_1)\,,$$

y si R_1 y B_1 son incompatibles, también lo son $R_2 \cap B_1$ y $R_2 \cap B_1$. Para obtener ahora $P(R_2)$ basta tener en cuenta la igualdad anterior y el axioma 3 de la definición de probabilidad:

$$P(R_2) = P(R_2 \cap R_1) + P(R_2 \cap B_1);$$

pero de la definición de probabilidad condicionada se deduce que $P(R_2 \cap R_1)=P(R_2|R_1)\ P(R_1)$ y $P(R_2 \cap B_1)=P(R_2|B_1)\ P(B_1)$, con lo que

$$P(R_2) = P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1) \qquad (1),$$

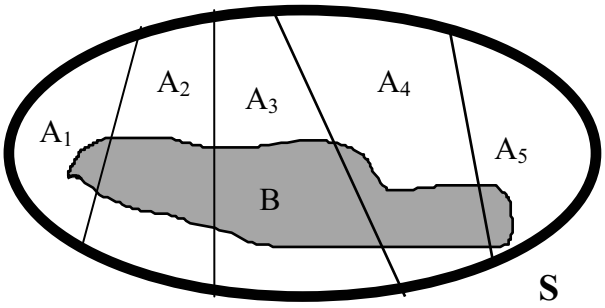
siendo todas ellas probabilidades que podemos obtener aplicando un razonamiento análogo al del ejemplo. En definitiva,

$$P(R_2) = \frac{19}{99}\cdot\frac{20}{100} + \frac{20}{99}\cdot\frac{80}{100} = \frac{20(19+80)}{99\cdot 100} = \frac{20\cdot 99}{99\cdot 100} = \frac{20}{100}.$$

Cuando dos o más sucesos verifican las condiciones de R_1 y B_1 decimos que constituyen una **partición** del espacio muestral. Si la partición está constituida por una familia $\{A_1, A_2, \ldots, A_n\}$ de sucesos, entonces la igualdad (1) adopta una forma más general que es conocida como el

Teorema de la Probabilidad Total Dado un suceso B y una partición, $\{A_1, A_2, \ldots, A_n\}$, del espacio muestral, con $P(A_i)>0$, para todo *i*, se verifica:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$



El esquema nos muestra una partición de S constituida por cinco sucesos, $\{A_1, A_2, A_3, A_4, A_5\}$, y el efecto que dicha partición tiene sobre el suceso B (en gris).

Ejemplo 7 (Una aplicación interesante del Teorema de la Probabilidad Total) Es bien conocida la reticencia de los encuestados a contestar a preguntas que pudiéramos calificar de *delicadas* (preguntas acerca de creencias religiosas, consumo de estupefacientes, hábitos sexuales, etc.), en la medida en que no se garantice el total anonimato en los cuestionarios. Los especialistas buscan métodos que tranquilicen, en ese sentido, a los encuestados. Veamos un método sencillo derivado de la aplicación del anterior teorema:

Se pretende conocer el consumo de drogas entre los 100 estudiantes de una clase. Para garantizar el anonimato colocamos 100 bolas numeradas del 1 al 100 en una urna y les hacemos extraer una bola cada uno, bola que conservan y no reponen a la urna. Aquellos cuyo número es menor o igual que 30 les pedimos que contesten, en un trozo de papel cualquiera, a la pregunta, *¿has consumido drogas alguna vez?*; los 70 restantes, que obtuvieron bolas numeradas del 31 al 100, deben de contestar a la pregunta, *¿termina tu DNI en una cifra par?* En ambos casos depositan el papel, adecuadamente doblado, en una urna dispuesta a tal efecto. Al proceder de esta forma solamente el interesado sabe a cual de las dos preguntas está contestando y por tanto el anonimato está plenamente garantizado.

Finalizado el proceso hemos contado los papeles que contienen un *SI*, y han resultado ser 42. Con esta información, ¿cuál es el número de estudiantes que han consumido drogas en alguna ocasión?

Si B={el número del estudiante es menor o igual que 30}, B^c={el número del estudiante es mayor que 30} y A={la respuesta ha sido SI}, aplicando el Teorema de la Probabilidad Total,

$$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C),$$

siendo P(A|B) la probabilidad que nos interesa calcular, puesto que aquellos cuyo número era menor o igual que 30 son los que respondieron a la pregunta que nos interesaba. Asimilando probabilidad a proporción, puesto que las condiciones lo permiten, sabemos

que P(A)=0,42, P(B)=0,30 y P(B^c)=0,70. Además, es razonable pensar que la mitad de los DNI terminarán en cifra par y la otra mitad en impar, de forma que aproximadamente P(A|B^c)=0,50. Sustituyendo en la igualdad anterior,

$$0{,}42 = P(A|B)\cdot 0{,}30 + 0{,}50\cdot 0{,}70 = 0{,}3\cdot P(A|B) + 0{,}35$$

y despejando,

$$P(A|B) = \frac{0{,}07}{0{,}3} = 0{,}233.$$

Luego, aproximadamente, 23 estudiantes han consumido drogas en alguna ocasión

3. Independencia de sucesos

En el ejemplo 5 hay un resultado que debiera sorprendernos. Hemos obtenido para P(H|C) el mismo valor que para P(H), lo que parece contradecir la idea de que la información recibida a priori debe modificar la probabilidad del suceso H. Cabe decir que la información en esta ocasión no ha servido para nada y hemos llegado a una conclusión que ya conocíamos. ¿Qué ha ocurrido?

Antes de responder comprobemos que esta es una situación que aparece en otros experimentos. Por ejemplo, consideremos la extracción de una carta al azar de un baraja española (48 cartas, 12 de cada palo). La probabilidad de *A={la carta extraída es un as}* es P(A)=4/48=1/12 y, si *B={la carta es de oros}*, tenemos que

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{1/48}{12/48} = \frac{1}{12},$$

que es el mismo resultado obtenido para el suceso sin condicionar.

La explicación reside en el hecho de la existencia de sucesos cuya probabilidad no se modifica por el conocimiento previo de la realización de otros. Una palabra que expresa adecuadamente este comportamiento es la de **independencia**. Observemos que si

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = P(A),$$

entonces

$$P(A\cap B) = P(A)P(B),$$

igualdad esta última que utilizaremos para definir el concepto de sucesos independientes.

Definición de sucesos independientes Dados dos sucesos cualesquiera, A y B, decimos que son *independientes* si:

$$P(A\cap B) = P(A)P(B)$$

Con esta definición de independencia de sucesos veamos lo que ocurre al tratar de obtener P(A|B),

$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

De la misma manera llegaríamos a P(B|A)=P(B), resultados ambos dos coherentes con la observación con la que iniciábamos el párrafo.

Ejemplo 8 ¿Qué ocurre en el ejemplo 6 si reemplazamos la bola extraída? Al obtener P(R_2 |R_1) vemos que las condiciones iniciales de la urna no se han modificado, hay 100 bolas y 20 de ellas son rojas, por lo tanto P(R_2 |R_1)=20/100 que coincide con P(R_2).

Así pues, cuando las extracciones se llevan a cabo con reemplazamiento, R_1, R_2 y en general, R_n (si hemos llevado a cabo n extracciones), son sucesos independientes.

Si llevamos a cabo 4 extracciones con reemplazamiento, a) ¿cuál es la probabilidad de obtener 1 bola roja? y, b) ¿de qué sean 3 las bolas rojas obtenidas?

a) Consideremos los sucesos

$A_1=\{R_1\cap B_2\cap B_3\cap B_4\}$, $A_2=\{B_1\cap R_2\cap B_3\cap B_4\}$, $A_3=\{B_1\cap B_2\cap R_3\cap B_4\}$ y $A_4=\{B_1\cap B_2\cap B_3\cap R_4\}$,

que representan, respectivamente, la bola roja aparece en la 1ª, 2ª, 3ª y 4ª extracción y son todos ellos incompatibles entre sí dada su definición. El suceso que nos interesa, A={se ha extraído una bola roja}=$A_1 \cup A_2 \cup A_3 \cup A_4$, por tanto

$$P(A)=P(A_1)+P(A_2)+P(A_3)+P(A_4).$$

Para obtener las probabilidades de los A_i hemos de tener en cuenta que las extracciones con reemplazamiento dan lugar a sucesos independientes, así

$$P(A_1) = P(R_1)P(B_2)P(B_3)P(B_4) = \frac{2}{10}\cdot\frac{8}{10}\cdot\frac{8}{10}\cdot\frac{8}{10} = \frac{2}{10}\cdot\left(\frac{8}{10}\right)^3,$$

resultado que merece ser comentado, porque ello nos evitará tener que calcular los restantes. En efecto, para obtener $P(A_1)$ sólo ha importado el número de bolas de cada color que hemos extraído, careciendo de importancia el lugar en que aparecieron. Siendo así, como todos los A_i comparten el mismo número de bolas rojas (1) y blancas (3), todos ellos tendrán igual probabilidad y podremos escribir,

$$P(A) = 4\cdot\frac{2}{10}\cdot\left(\frac{8}{10}\right)^3.$$

b) Cuanto hemos aprendido en el apartado anterior ha de facilitarnos la obtención de la probabilidad de B={se han extraído tres bolas rojas}. Comencemos por observar que hay una forma equivalente de definir B, diciendo que B={se ha extraído una bola blanca}. Análogamente a como hicimos antes, definimos ahora sucesos C_i , que representan, la bola blanca aparece en la i-ésima extracción y obtendremos para P(B),

$$P(B) = 4\cdot\frac{8}{10}\cdot\left(\frac{2}{10}\right)^3$$

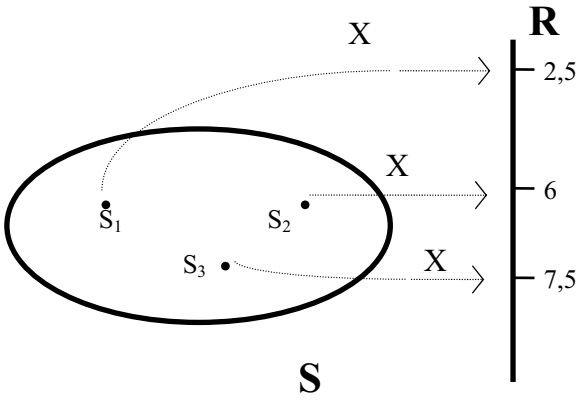
Una pregunta cuya respuesta daremos en el tema siguiente: si nos hubiesen dicho que las extracciones son 35, ¿podríamos soslayar el arduo problemas de contar una a una todas las distintas formas que tienen de aparecer las tres bolas rojas? Podemos adelantar que la respuesta es afirmativa.

TEMA 4.- VARIABLE ALEATORIA

1. Definición

Como hemos podido comprobar en alguno de los ejemplos de la Tabla 1 del tema anterior, el espacio muestral S es un espacio abstracto de difícil manejo. La mayor sencillez de manejo que los números nos ofrecen, y nuestra mejor capacidad para ello, explica que en alguno de aquellos ejemplos hayamos convertido el resultado en número, como en el caso del lanzamiento del dado, en el que hemos asignado a cada cara el número de puntos que contiene. De la misma manera, al ocuparnos del experimento que describe el sexo de los 3 hijos de un matrimonio, podemos hablar del número de varones o de mujeres.
Si hemos de trabajar con valores numéricos asociados a nuestro experimento, y nuestras limitaciones así nos lo aconsejan, conviene que lo hagamos correctamente mediante la definición del concepto de variable aleatoria:

Definición de variable aleatoria Una variable aleatoria, X, es una aplicación del espacio muestral S en el conjunto de los números reales, R, que a cada resultado le asigna un número real.



La variable aleatoria X asigna a los puntos s_1, s_2 y s_3 del espacio muestral S, los valores 2,5, 6 y 7,5, respectivamente

Dicho en lenguaje coloquial, se trata de asignar un número a cada resultado que obtengamos en nuestro experimento, y como éste es aleatorio nuestra variable también lo será. Es decir, sólo podremos conocer de ella con qué probabilidad tomará cada uno de sus posibles valores.

Es costumbre designar las variables aleatorias con las mayúsculas de las últimas letras del abecedario, X, Y, Z, y sus valores con las correspondientes minúsculas. Veamos algunos ejemplos de variables aleatorias definidas sobre espacios muestrales ya conocidos.

	Experimento	Espacio muestral	Variable(s) aleatoria(s)
E1	Lanzar dos monedas	S={CC,C+,+C,++}	X=número de caras en los lanzamientos
E3	Llamadas a una centralita telefónica	S={0,1,2,3,4,...}	X=número de llamadas
E4	Sexo de los 3 hijos de un matrimonio	S={VVV,VVM,VMV,MVV, VMM,MVM,MMV,MMM}	X=número de mujeres
E5	Elegir al azar un punto en un círculo de radio 1	S={Los puntos del círculo}	X=distancia del punto al centro Y=longitud de la cuerda que pasa por el punto
E7	Lanzar dos dados	S={(1,1),(1,2), …, (6,5),(6,6)}	X=suma de las caras Y=producto de las caras

Tabla 1.- **Algunos ejemplos de variables aleatorias**

La variable aleatoria como modelo probabilístico Las variables aleatorias no sólo nos permiten manejar con mayor facilidad los experimentos aleatorios, gracias a que convierten en valores numéricos sus resultados, sino que poseen otra cualidad que las hace todavía más interesantes y necesarias: su papel como **modelos probabilísticos** que describen un conjunto de experimentos que, aún siendo distintos, comparten características comunes.

¿De qué estamos hablando? Para comprenderlo, volvamos a la pregunta con la que iniciábamos el tema anterior. Queríamos conocer el tiempo que tardaríamos en recorrer los 350 kms. que separan Valencia de Barcelona, si nos movíamos a velocidad constante de 100 kms/hora. Supimos de inmediato que la respuesta era 3,5 horas y ello gracias a que conocemos la ecuación, $e=v\cdot t$. Esta ecuación es un **modelo determinista** que describe la relación entre espacio, velocidad y tiempo, y es válido para todos los móviles, cualquiera que sea el medio en el que se desplacen, a condición que lo hagan a velocidad constante, siendo ésta la característica común a todos ellos. Las variables aleatorias jugarán un papel semejante, pero en el contexto de los experimentos aleatorios.

Variables aleatorias y sucesos ¿Cómo calcular, por ejemplo, P(X=1) en E1? Observemos que cuando decimos que X=1 estamos diciendo que nos interesan todos aquellos resultados que conducen a que la variable tome el valor 1, pero esto es una forma de definir el suceso *A={ha salido una cara}*, luego A={X=1}={C+,+C}. Ahora podemos escribir que P(X=1)=P(A)=2/4. En definitiva, para obtener cualquier probabilidad relacionada con X obtendremos el correspondiente suceso y luego la probabilidad de éste

Así, si nos piden obtener P(X>8) en E7, obtendremos en primer lugar aquel suceso cuyos puntos hacen que la suma de ambas caras sea superior a 8, a saber

$$\{X>8\}=\{(3,6),(6,3),(4,5),(5,4),(4,6),(6,4),(5,5),(5,6),(6,5),(6,6)\}$$

y de aquí P(X>8)=10/36. Si en el mismo experimento se nos pide obtener $P(X\leq 3, Y=2)=P(\{X\leq 3\}\cap\{Y=2\})$, habremos de obtener el suceso que contiene aquellos resultados cuya suma de caras no supere a 3 y que, simultáneamente, su producto valga 2. Estamos hablando del suceso A={(1,2),(2,1)} y por lo tanto $P(X\leq 3, Y=2)=2/36$.

Ejemplo 1 Tenemos un urna con 100 bolas, 30 de ellas son blancas, 50 son azules y las 20 restantes son rojas. Llevamos a cabo dos extracciones con reemplazamiento y definimos las variables

X=número de bolas blancas
Y=número de bolas azules, y
Z=número de bolas rojas.

¿Qué valores pueden tomar cada una de estas variables? ¿Qué valen P(X=1), P(Z=2), P(Y=0) y P(X=1, Z=2)?

En primer lugar observemos que el espacio muestral viene dado por,

$$S=\{B_1B_2, B_1A_2, B_1R_2, A_1B_2, A_1A_2, A_1R_2, R_1B_2, R_1A_2, R_1R_2\},$$

en donde la letra designa el color y el subíndice la extracción. Como las extracciones se llevan a cabo con reemplazamiento, la obtención de cualquier color es independiente de una extracción a otra. Además, por lo visto en el ejemplo 8 del tema anterior, la probabilidad de obtener un determinado color es la misma en todas las extracciones. Lo que supone que

$$P(B_1)=P(B_2)=3/10,\quad P(A_1)=P(A_2)=5/10,\quad P(R_1)=P(R_2)=2/10.$$

Del espacio muestral anterior deducimos que las tres variables pueden tomar los mismos valores y estos son: 0, 1, 2. Para obtener las probabilidades que se nos piden, vamos a determinar con que sucesos se corresponde cada una de ellas:

$\{X=1\}=\{B_1A_2, B_1R_2, A_1B_2, R_1B_2\}$,
$\{Z=2\}=\{R_1R_2\}$,
$\{Y=0\}=\{B_1B_2, B_1R_2, R_1R_2, R_1B_2\}$,
$\{X=1\}\cap\{Z=2\}=\varnothing$

Las probabilidades buscadas son:

$$P(X=1)=\frac{3}{10}\cdot\frac{5}{10}+\frac{3}{10}\cdot\frac{2}{10}+\frac{5}{10}\cdot\frac{3}{10}+\frac{2}{10}\cdot\frac{3}{10}=\frac{42}{100}$$

$$P(Z=2)=\frac{2}{10}\cdot\frac{2}{10}=\frac{4}{100}$$
$$P(Y=0)=\frac{3}{10}\cdot\frac{3}{10}+\frac{3}{10}\cdot\frac{2}{10}+\frac{2}{10}\cdot\frac{2}{10}+\frac{2}{10}\cdot\frac{3}{10}=\frac{25}{100}$$
$$P(X=1,Z=2)=0$$

Tipos de variables aleatorias Atendiendo a los valores que pueden tomar, las variables aleatorias se clasifican en:

- **discretas:** cuando toman valores aislados, en número finito o infinito. En la Tabla 1 podemos encontrar variables discretas de ambos tipos: la variable X del experimento E1 es discreta con un número finito de valores, mientras que la variable X del experimento E3 es discreta pero puede tomar infinitos valores.
- **continuas:** las que toman cualquiera de los valores comprendidos entre los extremos de un intervalo, que puede tener uno o sus dos extremos infinitos. En la Tabla 1, la variable X del experimento E5 toma valores en el intervalo [0,1], puesto que podemos encontrar puntos cuya distancia al centro es cualquier valor comprendido entre 0 y 1.

El siguiente esquema recoge la clasificación de las variables aleatorias:

$$\textit{variables aleatorias}\begin{cases}\textit{discretas}\begin{cases}\textit{finitas}\\ \textit{infinitas}\,(\textit{numerables})\end{cases}\\ \textit{continuas}\begin{cases}\textit{acotadas}\\ \textit{no acotadas}\end{cases}\end{cases}$$

Media y varianza de una variable aleatoria discreta Al igual que hicimos con las variables observadas en una muestra, podemos también ahora definir medidas de posición y localización para las variables aleatorias. Las más utilizadas son la **media** y la **varianza** y la raíz cuadrada de ésta, la **desviación típica**. Para distinguirlas de las correspondientes medidas calculadas para variables observadas en una muestra, se las designa mediante letras del alfabeto griego. Para una variable aleatoria discreta, $X=\{x_1, x_2, x_3,\ldots, x_n\}$,

- **la media**, μ, se obtiene de la siguiente forma:

$$\mu = x_1\cdot P(X=x_1) + x_2\cdot P(X=x_2) + \ldots + x_n\cdot P(X=x_n) = \sum_{i=1}^{n} x_i P(X=x_i)\,.$$

- **la varianza**, σ^2, se obtiene de la siguiente forma:

$$\mu = (x_1-\mu)^2\cdot P(X=x_1) + (x_2-\mu)^2\cdot P(X=x_2) + \ldots + (x_n-\mu)^2\cdot P(X=x_n) =$$
$$= \sum_{i=1}^{n} (x_i-\mu)^2 P(X=x_i)\,,$$

y no es más que la media de los cuadrados de las desviaciones de cada valor de la variable respecto de su media.

Ejemplo 2 Si al lanzar un dado definimos la variable X=*número de puntos que nos muestra la cara del dado*, tendremos que $X=\{1,2,3,4,5,6\}$ y como se dan las condiciones para aplicar la fórmula de la Laplace, la variable tomará cualquiera de estos valores con la misma probabilidad. Por ejemplo, para X=1,

$$P(X=1)=\frac{\text{casos favorables}}{\text{casos posibles}}=\frac{1}{6}\,,$$

y su media valdrá,

$$\mu = 1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6}=\frac{1+2+3+4+5+6}{6}=\frac{21}{6},$$

que coincide con la media aritmética de los valores que toma nuestra variable aleatoria, como no podía ser de otra forma si recordamos la analogía existente entre probabilidad y frecuencia relativa.
La varianza valdrá,

$$\sigma^2=\left(1-\frac{21}{6}\right)^2\cdot\frac{1}{6}+\left(2-\frac{21}{6}\right)^2\cdot\frac{1}{6}+\left(3-\frac{21}{6}\right)^2\cdot\frac{1}{6}+\left(4-\frac{21}{6}\right)^2\cdot\frac{1}{6}+$$
$$+\left(5-\frac{21}{6}\right)^2\cdot\frac{1}{6}+\left(6-\frac{21}{6}\right)^2\cdot\frac{1}{6}=\frac{105}{36}.$$

La desviación típica será la raíz cuadrada del anterior valor

$$\sigma=\frac{\sqrt{105}}{6}.$$

Puesto que los experimentos aleatorios pueden tener orígenes muy distintos, son muchos los posibles modelos probabilísticos necesarios para describirlos. Dado el carácter introductorio de este curso, nos limitaremos a estudiar y aplicar dos de los modelos más conocidos: el **modelo binomial**, para variables discretas finitas, y el **modelo normal**, para variables continuas.

2. La distribución Binomial

En el ejemplo 8 del tema anterior nos ocupábamos de un experimento consistente en extraer, con reemplazamiento, 4 bolas de una bolsa que contenía 20 bolas rojas y 80 blancas. Obtuvimos la probabilidad de los sucesos, *A={1 bola roja entre las cuatro extraídas}*, y, *B={3 bolas rojas entre las cuatro extraídas}*. Si definimos la variable, X=*número de bolas rojas entre las cuatro extraídas*, cuyos posibles valores son {0,1,2,3,4}, aquellas probabilidades se corresponden con P(X=1) y P(X=3), respectivamente. Recordemos que valían:

$$P(X=1)=4\cdot\frac{2}{10}\cdot\left(\frac{8}{10}\right)^3 \quad \text{y} \quad P(X=3)=4\cdot\frac{8}{10}\cdot\left(\frac{2}{10}\right)^3.$$

Supongamos ahora que la composición de la bolsa cambia y que hay 50 bolas de cada color. En este nuevo experimento, como la probabilidad de obtener bola roja en cada extracción es 1/2 en lugar de 2/10, y la correspondiente para la bola blanca es 1/2 en lugar de 8/10, las anteriores probabilidades valdrán:

$$P(X=1)=4\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3 \quad \text{y} \quad P(X=3)=4\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3$$

Consideremos ahora dos nuevos experimentos y las correspondientes variables:

- lanzamiento de cuatro monedas, Y=número de caras en los cuatro lanzamientos
- sexo de los cuatro hijos de un matrimonio, Z=número de mujeres entre los cuatro hijos,

en los que queremos conocer P(Y=1) y P(Y=3) y P(Z=1) y P(Z=3), respectivamente. Podemos, para ello, comenzar a razonar como lo hicimos antes, pero ¿vale la pena hacerlo, a la vista del resultado anterior? o si se prefiere, ¿podemos aprovechar aquel resultado para dar una respuesta? Para responder a estas preguntas nos conviene primero tratar de dilucidar qué aspectos en común tienen, si los tienen, nuestros tres experimentos:

1. los tres consisten en el mismo número de repeticiones, *cuatro*, de pruebas que son independientes entre sí, en la medida que conducen a sucesos que lo son. Estas pruebas son, respectivamente,
    a) cuatro extracciones de bolas con reemplazamiento,
    b) lanzamientos de cuatro monedas, que equivale a cuatro lanzamientos de una misma moneda, y
    c) cuatro nacimientos de los hijos de una misma familia,
2. en todos ellos nos interesamos por la *ocurrencia de un suceso*, A, el mismo en cada una de las cuatro pruebas; contabilizando el número total de ocurrencias del suceso A en las

cuatro pruebas mediante la correspondiente variable aleatoria. Se trata, en cada caso, del suceso

a) A={la bola extraída es roja},
b) A={la moneda muestra una cara}, y
c) A={el sexo del niño es mujer},

3. la probabilidad del suceso que nos interesa permanece constante en las cuatro pruebas y, en particular, para los tres experimentos es la misma y vale 1/2.

Todo esto supone que, con independencia de que hablemos de bolas, monedas o nacimientos, los resultados obtenidos para las extracciones de bolas son aplicables a las restantes situaciones y

$$P(X=1)=P(Y=1)=P(Z=1)=4\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3$$

$$P(X=3)=P(Y=3)=P(Z=3)=4\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)^3$$

Todos los experimentos que comparten las tres características antes señaladas pueden ser descritos mediante la variable aleatoria **Binomial**, también conocida como la **distribución Binomial**, cuya definición formal damos a continuación.

Definición de variable aleatoria Binomial Supongamos que llevamos a cabo un experimento aleatorio con las siguientes características:

1. se llevan a cabo **n** pruebas independientes, todas ellas en las mismas condiciones,
2. en cada prueba nos interesamos en un mismo suceso, A, a cuya ocurrencia denominaremos **éxito**, y
3. la probabilidad de que ocurra un éxito es la misma en cada prueba, **p.**

La variable aleatoria X=*número de éxitos en la* n *pruebas*, recibe el nombre de **Binomial** con **parámetros n** y **p** (se dice también, que sigue la **distribución Binomial)**. La probabilidad de que hayan ocurrido *k* éxitos en las *n* pruebas, P(X=k), se obtiene mediante la fórmula:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,\cdots,n$$

donde el número combinatorio $\binom{n}{k}=\frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots2\cdot1}$, representa las distintas formas en que pueden aparecer los *k* éxitos en las *n* pruebas.

Es costumbre designar abreviadamente a la variable aleatoria Binomial mediante **B(n,p)**. A los valores **n** y **p** se les denomina parámetros porque ellos son los que caracterizan la distribución y distinguen una Binomial de otra y también porque, una vez conocidos, podemos obtener cualquier probabilidad relacionada con X. Obsérvese que X puede tomar *n+1* valores, concretamente {0,1,2,3,…,n}.

Media, varianza y desviación típica de una B(n,p) Para una variable aleatoria Binomial de parámetros *n* y *p*, se tiene,

$$\mu=np,\qquad \sigma^2=np(1-p),\qquad \sigma=\sqrt{np(1-p)}.$$

Ejemplo 3 Volviendo a los experimentos con los que abríamos este párrafo, las tres variables involucradas son todas ellas una misma Binomial, concretamente B(4,1/2). Si quisiéramos obtener la probabilidad de que nuestras cuatro extracciones contengan 2 bolas rojas,

$$P(X=2)=\binom{4}{2}\left(\frac{1}{2}\right)^2\cdot(1-\frac{1}{2})^2=\frac{4\cdot3}{2\cdot1}\left(\frac{1}{2}\right)^4$$

La media de X vale $\mu=2$, su varianza $\sigma^2=1$ y su desviación típica $\sigma=1$.

Tablas de la Binomial Los valores de P(X=k) para una variable B(n,p) vienen tabulados en tablas específicas, lo que facilita su obtención y evita cálculos engorrosos. En el Anexo I aparece una de estas tablas, cuyo contenido y manejo explicaremos a continuación.

La tabla está encabezada por una **primera fila** en la que aparecen los distintos valores de *p*. En la **primera columna** aparecen los distintos valores de *n* que recoge la tabla (las tablas son tanto mejores cuantos más valores de *p* y *n* recogen), y en la **segunda columna**, encabezada por la letra *k*, figuran los n+1 valores que toma la variable para cada *n*. Cada valor del interior de la tabla se corresponde con una terna única de valores de *n*, *p* y *k* y representa el valor de P(X=k). Ilustremos su uso mediante el siguiente ejemplo:

Ejemplo 4 Sabemos que, entre la población de la que se tomó la muestra de 250 personas recogida en le Anexo I del tema I, existen un 20% de fumadores. Si extraemos al azar una muestra de 10 personas, ¿cuál es la probabilidad de que nos encontremos con 3 fumadores?, ¿y la de que el número de fumadores no supere a dos?
Aunque no se dice nada acerca del tipo de extracción, en este tipo de muestreo no suele haber reemplazamiento, pero como el tamaño de la población de la que se extrae la muestra es mucho mayor que el tamaño de la muestra, la proporción de fumadores permanece prácticamente inalterada después de cada extracción y podemos suponer que hay independencia de una extracción a otra. En este contexto, el número de fumadores en la muestra de 10 personas será una B(10,0.2). Para obtener las probabilidades que se nos piden haremos uso de la tabla, reproduciendo aquí la parte que nos interesa.

				p		
n	*k*	…	1/6	.20	.25	…
10	*0*	.	.1615	.1074	.0563	.
	1	.	.3230	.2684	.1877	.
	2	.	.2907	.3020	.2816	.
	3	.	.1550	.2013	.2503	.
	4	.	.0543	.0881	.1460	.
	5	.	.0130	.0264	.0584	.
	6	.	.0022	.0055	.0162	.
	7	.	.0002	.0008	.0031	.
	8	.	.0000	.0001	.0004	.
	9	.	.0000	.0000	.0000	.
	10	.	.0000	.0000	.0000	.
11	*0*	.	.1346	.0859	.0422	.
.	.	.	…	…	…	.

En la primera fila buscamos la columna encabezada por *p=0.2*, descendiendo por ella hasta la fila correspondiente a *k=3*, *n=10*, tal como se muestra en el esquema. En la confluencia de la fila y columna encontramos P(X=3) = 0.2013.
La segunda probabilidad que se nos solicita es $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$, y procediendo como antes tendremos que

$$P(X \leq 2) = 0.1074 + 0.2684 + 0.3020 = 0.6778.$$

3. La distribución Normal

Cuando estudiamos cualquier variable en la muestra (Tema 1), debemos ser conscientes que estamos analizando un conjunto de datos que provienen de una **población**, de la cual han sido extraídos mediante un muestreo aleatorio. Para el caso de las variables continuas, el peso y la altura, por ejemplo, el experimento ha consistido en extraer una persona al azar y proceder a pesarla o tallarla, siendo el resultado de dichos procesos de medida sendas variables aleatorias de las que la muestra nos proporciona un conocimiento parcial. El comportamiento probabilístico de la variable peso, o de cualquier otra variable continua, es conocido a través de lo que denominábamos su **curva de frecuencias** o **curva de densidad**, que en el tema 1 obtuvimos haciendo crecer el tamaño de la muestra de pesos.

Cada variable aleatoria tiene su propia curva de densidad, así, si estudiamos ahora las alturas y procedemos como hicimos allí con los pesos, obtendremos la curva de densidad que aparece en la Figura 1 y cuya forma es semejante a la que obtuvimos para el peso, aunque con escalas distintas puesto que se trata de magnitudes distintas. Esta semejanza es debida a que ambas variables aleatorias se comportan de la misma manera y, en consecuencia, comparten el mismo modelo probabilístico: la llamada **distribución normal** y decimos que ambas son **variables aleatorias normales**.

Conviene ahora recordar, como señalábamos cuando la introducíamos, que la curva de densidad nos proporciona las frecuencias relativas mediante áreas. Así, la frecuencia relativa de los valores de las alturas comprendidos entre 150 y 188 es el área encerrada bajo la curva entre dichos valores, que en la figura 1 aparece sombreada. Ello supone que el área total encerrada bajo la curva vale la **unidad**, el total de las frecuencias relativas.

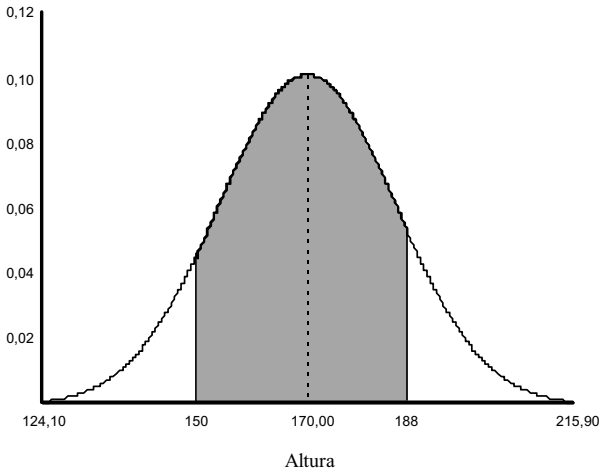


Figura 1.- **Curva de densidad correspondiente a la altura**

Forma y parámetros de la distribución normal La forma de la distribución normal, o para ser más precisos de su **curva de densidad**, es la de una campana simétrica, razón por la cual se la conoce también con el nombre de **campana de Gauss**, en honor al primer matemático que se ocupó de ella. Pero dentro de esta forma general, la curva correspondiente a cada variable aleatoria normal presenta peculiaridades que dependen de sus parámetros: **media** y **varianza**.

- La media, μ, en tanto que medida de localización, nos permite situar la curva sobre la escala de valores de la variable y, en particular, indica la posición del **eje de simetría** de la curva.
- La varianza, σ^2, como medida de dispersión que es, nos indica la mayor o menor dispersión de los valores de la variable alrededor de su media. A mayor varianza, mayor dispersión, y la curva es más abierta y con menor altura, mientras que una varianza menor, al implicar menor dispersión, da lugar a curvas más elevadas y más cerradas, puesto que el área que encierra la curva es siempre la misma e igual a uno.

Puesto que media y varianza caracterizan la distribución normal, cuando una variable aleatoria es normal siempre debemos especificar su media y su varianza, lo que se hace habitualmente escribiendo: **X es N(μ,σ^2)**.

La figura 2 muestra la superposición de tres curvas normales correspondientes a otras tantas variables que comparten la misma media, $\mu = 170$, y cuyas varianzas son distintas, $\sigma_1^2 = 234.09$, $\sigma_2^2 = 58.52$ y $\sigma_3^2 = 14.63$.

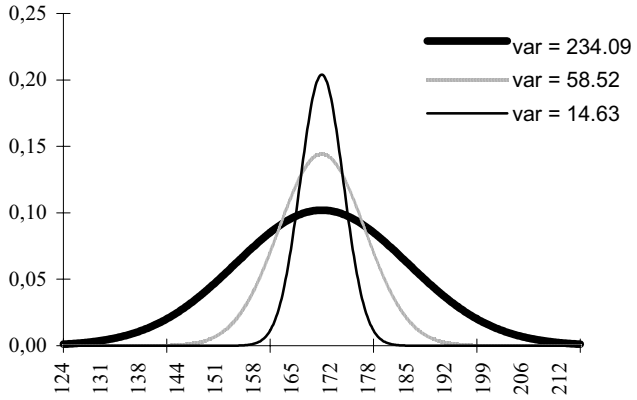


Figura 2.- **Curvas de densidad normales con igual media y distintas varianzas**

Obtención de probabilidades para la distribución normal Como ya hemos dicho, si X es una variable aleatoria normal, la probabilidad de que esté comprendida entre dos valores, a y b, $P(a \le X \le b)$, viene dada por el área que encierra la curva entre ambos valores, tal como se muestra en la figura 1. Eventualmente, alguno de los dos valores puede ser infinito como, por ejemplo, cuando nos piden la probabilidad de que X no supere el valor b, $P(X \le b)$, que viene representada por el área que encierra la curva a la izquierda de b.

Por otra parte, como área y probabilidad son equivalentes ahora, todas las probabilidades que escribimos a continuación son iguales, por cuanto el añadir o quitar un punto, no altera el área del recinto:

$$P(a < X \le b) = P(a < X < b) = P(a \le X \le b) = P(a \le X < b).$$

Pero ¿cómo calculamos las áreas bajo la curva normal? Al igual que ocurría con la distribución Binomial, también ahora contamos con la ayuda de unas tablas que nos facilitan el cálculo, si bien es necesario llevar a cabo previamente una transformación de nuestros datos por un motivo que fácilmente comprenderemos.

La normal tipificada Ya hemos visto que las características que definen una distribución normal son su media y su varianza, lo que supone que hemos de disponer de tablas para todas las posibles medias y varianzas con la que nos podamos encontrar. Pero si tenemos en cuenta que la media de una normal puede tomar cualquier valor positivo o negativo, y que su varianza puede ser cualquier valor positivo, llegaremos a la conclusión de que resulta imposible poder abarcar tal diversidad de valores.

En realidad tampoco es necesario, porque con una sola tabla nos bastará para poder obtener las probabilidades de cualquier distribución normal. En efecto, mediante la transformación que denominábamos **tipificación**, cualquier variable X, $N(\mu,\sigma^2)$, se convierte en la variable Z, N(0,1), conocida ésta como la **normal tipificada**. Recordemos que la tipificación consistía en,

$$Z = \frac{X - \mu}{\sigma}.$$

¿Cómo hacer uso de esta transformación? Supongamos que queremos conocer la probabilidad de que la altura de una persona extraída al azar está comprendida entre 150 y 188 cm., sabiendo que dicha variable, X, es N(170,234.09). Si aplicamos la transformación tendremos,

$$X = 150 \text{ da lugar a } Z = \frac{150 - 170}{\sqrt{234,09}} = -1,3071$$

$$X = 188 \text{ da lugar a } Z = \frac{188 - 170}{\sqrt{234,09}} = 1,1763$$

y decir que *X está comprendida entre 150 y 188 cm.* equivale a decir que *Z está comprendida entre -1,3071 y 1,1763*. Ahora debemos aprender a buscar en la tabla de la N(0,1) los valores que les corresponden.

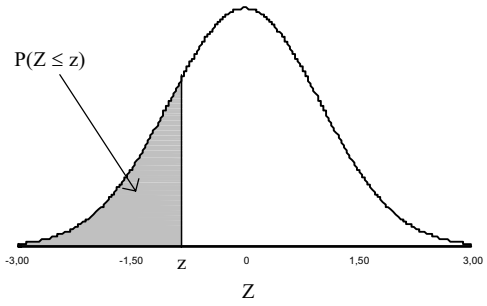


Figura 3.- **Interpretación de los valores de la tabla de la N(0,1)**

Uso de la tabla de la N(0,1) Se trata de una tabla de doble entrada (ver Anexo II) que nos proporciona, para los valores de Z desde -3 a 3, de centésima en centésima, el área encerrada bajo la curva a la izquierda del valor *z* elegido; es decir, $P(Z \leq z)$, tal como aparece en la figura 3. En el margen izquierdo de la tabla aparecen los valores de Z de décima en décima, debiendo desplazarnos a lo largo de la fila hasta encontrar la columna encabezada por la cifra correspondiente a la centésima deseada. Así, si queremos conocer el área que corresponde al valor $z = -1,75$, entraremos en la tabla por la fila cuyo valor es -1,7, desplazándonos hasta la sexta columna, la encabezada por un 5, y encontraremos el valor 0,0401. Así pues, $P(Z \leq -1,75) = 0,0401$.

Ejemplo 5 Extraída una persona al azar y sabiendo que su altura, X, es una N(170,234.09),

a) ¿cuál es la probabilidad de que su altura no supere los 165 cm.?,
b) ¿cuál es la probabilidad de que su altura esté comprendida entre 150 y 188 cm.?, y
c) ¿cuál es la probabilidad de que la persona supere los 180 cm.?

En el apartado a) se nos pide $P(X \leq 165)$ y llevaremos a cabo la tipificación correspondiente,

$$X = 165 \text{ da lugar a } Z = \frac{165 - 170}{\sqrt{234,09}} \approx -0,33$$

y tendremos que $P(X \leq 165) = P(Z \leq -0,33) = 0,3707$.

En el apartado b) se nos pide $P(155 \leq X \leq 188) = P(-1,3071 \leq Z \leq 1,1763)$, pero como en la tabla sólo podemos encontrar probabilidades del tipo $P(Z \leq z)$, observemos que

$$P(-1,3071 \leq Z \leq 1,1763) = P(Z \leq 1,1763) - P(Z \leq -1,3071),$$

y como la precisión de la tabla es de centésimas, buscaremos los valores más próximos a 1,1763 y -1,3071, que son 1,18 y -1.31. En definitiva,

$$P(155 \leq X \leq 188) = P(-1,3071 \leq Z \leq 1,1763) = 0,8810 - 0,0951 = 0,7859.$$

En el apartado c) se nos pide $P(X > 180)$, pero observemos que los sucesos $\{X > 180\}$ y $\{X \leq 180\}$ son complementarios, por tanto $P(X \leq 180) + P(X > 180) = 1$ y de aquí,

$$P(X > 180) = 1 - P(X \leq 180) = 1 - P(Z \leq 0,65) = 1 - 0,7422 = 0,2578,$$

donde 0,65 es el valor resultante de transformar 180 mediante la tipificación.

4. Aproximación de la Binomial mediante la Normal

El máximo valor de *n* que podemos encontrar en la tabla de la Binomial es 12. Incluso en tablas más completas este valor máximo no suele sobrepasar 50. ¿Cómo obtener entonces probabilidades cuando *n* supere estos valores? Un conocido resultado del Cálculo de Probabilidades nos permite soslayar el problema. Este resultado afirma que,

> si X es una variable aleatoria B(n,p), para valores de *n* suficientemente grandes, se comporta como si fuera una N(np, np(1-p)), es decir, como una normal con media, μ=np, y varianza, σ^2=np(1-p).

Dos preguntas surgen de inmediato:

1. ¿qué se entiende por *suficientemente grande*?,
2. ¿cómo utilizar en la práctica este resultado?

Como respuesta a la primera, digamos que la aproximación es buena cuando *n* y *p* son tales que *np>5* y *n(1-p)>5*. La mejor respuesta a la segunda pregunta es un ejemplo de aplicación como el que presentamos a continuación.

Ejemplo 6 La proporción de fumadores en determinada población es p=0,15. Si extraemos al azar una muestra de 40 personas y designamos por X el número de fumadores en la muestra,

a) ¿cuál es la probabilidad de que X≤8?,
b) ¿cuál es la probabilidad de que 3≤X≤10?,

Como *np=40·0,15=6*, y *n(1-p)=40·0,85=34*, ambos mayores que 5, podemos aplicar la aproximación y suponer que X es N(np,np(1-p)), es decir, X es aproximadamente N(6,5.1). Ahora actuaremos como si de una distribución normal se tratara y obtendremos las probabilidades requeridas recurriendo a la tipificación, para así poder utilizar la tabla de la N(0,1).

En el apartado a) se nos pide P(X≤8) y por tanto

$$P(X \le 8) = P\left(\frac{X-6}{\sqrt{5,1}} \le \frac{8-6}{\sqrt{5,1}}\right) \approx P(Z \le 0,89) = 0,8133\,.$$

En el apartado b) se nos pide P(3≤X≤10), que después de tipificar se puede aproximar por

$$P(3 \le X \le 10) \approx P(-1,33 \le Z \le 1,77) = P(Z \le 1,77) - P(Z \le -1,33) = 0,8698\,.$$

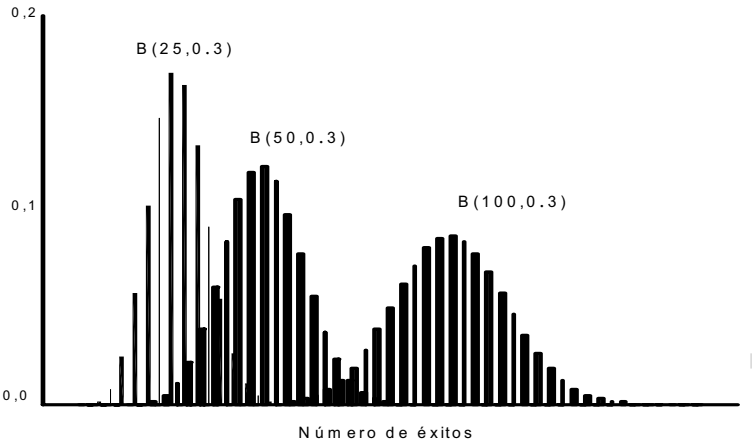


Figura 4.- **Aproximación de la B(n,p) a una Normal a medida que aumenta *n***

TEMA 5.- MUESTREO ALEATORIO

1. Muestra y Población

Un grupo de criminólogos ha desarrollado un test para predecir futuros comportamientos delincuentes de los adolescentes. El test consiste en una serie de preguntas cuyas respuestas son puntuadas, dando lugar a una valoración global del test. Las pruebas que se han realizado hasta el momento son prometedoras por cuanto, aplicado a delincuentes y no delincuentes, los resultados en ambos grupos son claramente diferentes, siendo las puntuaciones mayores en el primer grupo. Estos resultados han permitido también establecer que la variable aleatoria, X, valoración global obtenida en el test, sigue una distribución normal. Para que el test pueda ser utilizado con la finalidad para la que fue diseñado, necesitamos conocer cuales son los parámetros de la correspondientes distribución normal, es decir, su media, μ, y su varianza, σ^2.

En Economía, en Sociología, en Ciencias Experimentales, en Ciencias de la Salud, en procesos de fabricación, … y en prácticamente cualquier actividad humana, se nos presentan situaciones como la descrita en el párrafo anterior, en la que se desea conocer el comportamiento de una variable aleatoria en una **población**. Pero además todas ellas comparten el mismo problema: la imposibilidad de acceder a toda la población involucrada para poder así conocer el comportamiento de la variable, o característica poblacional, objeto de nuestro interés.

Una solución parcial al problema consiste en acceder a una parte de la población, **muestra**, y estudiar en ella las características que nos interesan. La **Inferencia Estadística** nos proporcionará los *procedimientos* para analizar los datos de la muestra y las *reglas* para extender nuestras conclusiones a la población de la cual procede la muestra.

Estimar y contrastar Dos son los procesos fundamentales que la Inferencia Estadística nos permitirá llevar a cabo:

1. **Estimar** las características poblacionales que nos interesan, o
2. **Contrastar** las hipótesis que acerca de dichas características hayamos conjeturado.

Aún cuando las posibles características poblacionales puedan ser muchas, en este curso nos vamos a ocupar solamente de cuanto concierne a las más habituales: *proporciones*, *medias* y *varianzas*.

Tabla de números aleatorios Puesto que el único acceso posible a la población va a ser a través de una muestra, conviene recordar lo que dijimos en el tema 2 acerca del muestreo aleatorio, puesto que si queremos que nuestra muestra sea representativa habrá de ser adquirida mediante un muestreo de estas características, a saber:

1. todos los miembros de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra, y
2. los elementos de la muestra son elegidos independientemente unos de otros, es decir, que la presencia de cualquiera de ellos en la muestra no influye en la futura presencia de los demás.

Pero sabido esto, cabe preguntarse acerca de los mecanismos de elección que garanticen que ambas condiciones se cumplen. Existen muchos y el desarrollo de la informática ha facilitado el acceso a los mismos. Aunque también ahora podemos contar con el auxilio de una tabla, de gran ayuda cuando de elegir muestras de tamaño pequeño se trata. Una **tabla de números aleatorios**, que así es conocida, es reproducida en el anexo y contiene 10,000 dígitos aleatoriamente generados (existen tablas que contienen hasta 1,000,000 de dígitos). Para mayor comodidad los dígitos aparecen agrupados de 5 en 5, aunque fueron generados individual e independientemente unos de otros. Veamos mediante un ejemplo cómo utilizarla:

Ejemplo 1 Queremos extraer una muestra de tamaño 10 de una población formada por 10,000 individuos. Previamente habremos numerado los individuos de 0 a 9,999 y a continuación nos situaremos en un punto cualquiera de la tabla; por ejemplo señalando al azar con la punta de un lápiz, o simplemente eligiendo una fila y una columna. Supongamos que nos hemos situado en la fila 31, columna 26, a partir de aquí, desplazándonos hacia la derecha, iremos leyendo los números resultantes de agrupar los dígitos de 4 en 4, deteniendo el proceso cuando tengamos los 10 números requeridos. En nuestro caso obtendremos:

4761 6557 5137 1873 1413 3113 2640 9392 9482 1565

La muestra estará formada por los elementos de la población cuyo número coincida con uno de los anteriores.

La elección del dígito inicial, a partir del cual iniciaremos la búsqueda, carece de importancia si vamos a elegir una sola muestra, pero si son varias conviene utilizar algún mecanismo que garantice una cierta aleatoriedad en el punto de arranque, o al menos que evite que siempre sea el mismo.

Ejemplo 2 Si queremos obtener una muestra de 10 individuos de los 250 que hemos utilizado en el anexo del tema 1, procederemos como hemos indicado antes, haciendo los cambios necesarios para adaptarnos a la nueva población. A saber, numerar de 0 a 249 y elegir ahora los dígitos de 3 en 3; pero aún con estos cambios podemos obtener un número mayor que 249. ¿Qué hacer? Sencillamente ignorarlo y continuar hasta que hayamos conseguido 10 números válidos, es decir, menores o iguales que 249.

Por ejemplo, si el punto de arranque es ahora el cruce de la fila 17 y la columna 49, tendríamos:

926 789 **143** **187** 581 592 414 429 683 **042** 760 298 704 571 **183** 340 429 **152**
487 394 999 733 **040** **045** 473 **049** 852 800 422 826 938 754 773 525 675 508
210 273 **027**

Elegida la muestra estudiaremos en ella la o las características que nos interesan. Por ejemplo, en la anterior muestra los 10 individuos queremos estudiar su sexo, su opinión sobre los jueces y su altura, como forma de conocer el comportamiento de estas variables en la población original. Más concretamente, queremos estimar la proporción de varones, la proporción de individuos que tiene buena opinión de los jueces y la media y varianza de la distribución de alturas. La proporción de elementos de una muestra que verifican cierta condición, suele representarse por $\hat{\pi}$, y la correspondiente proporción en la población de la que la muestra fue extraída, por p (o π).

Los datos obtenidos en la muestra son:

individuo	**sexo**	**opinión jueces**	**altura**
027	v	indiferente	167
040	v	indiferente	176
042	m	ns/nc	138
045	m	mala	169
049	v	mala	173
143	v	mala	180
152	v	indiferente	179
183	m	buena	174
187	v	buena	165
210	m	buena	167

En esta muestra las características que nos interesan valen:

$$\hat{\pi}_{varon} = 0{,}6 \qquad \hat{\pi}_{buena\ op} = 0{,}3 \qquad \bar{x}_{altura} = 168{,}8 \qquad s^2_{altura} = 143{,}95$$

y podríamos utilizarlas como estimaciones de los correspondientes valores en la población. Pero si ahora repetimos el proceso y obtenemos nuevas muestras, aun siendo éstas aleatorias y, por tanto, *representativas* de la población de la que han sido extraídas, pueden dar lugar a valores distintos.

En la tabla siguiente, que recoge los valores obtenidos para ésta y otras 9 muestras, podemos comprobar estos cambios:

muestra	$\hat{\pi}_{varon}$	$\hat{\pi}_{buena\ op}$	$\bar{x}_{altura}$	s^2_{altura}
01	0,6	0,3	168,80	143,95
02	0,5	0,6	175,48	150,87
03	0,5	0,4	159,45	190,65
04	0,3	0,7	169,57	140,13
05	0,4	0,3	174,35	187,13
06	0,2	0,3	172,38	145,67
07	0,4	0,5	170,03	198,15
08	0,6	0,4	169,26	159,58
09	0,5	0,3	165,78	190,45
10	0,4	0,5	169,95	158,49

Esta variabilidad parece dificultar cualquier decisión acerca de los valores que las características observadas (proporción, media y varianza) toman en la población. Es cierta la dificultad, pero la variabilidad observada era previsible por cuanto estamos estudiando variables que son aleatorias y, además, lo estamos haciendo a través de una muestra. Existe una doble fuente de variación aleatoria: la debida al comportamiento aleatorio de la variable en estudio y la que tiene su origen en el mecanismo aleatorio de elección de la muestra. La pregunta es cómo soslayar el problema, si es posible hacerlo, porque la única forma de acceder a la población es a través de una muestra aleatoria de la misma.

La Inferencia Estadística, como ya adelantábamos al principio, nos proporciona las herramientas para abordar y resolver el problema, pero la condición previa es conocer el comportamiento de las características que estudiamos en una muestra, pues de acuerdo con el comentario del párrafo anterior los valores que toman varían aleatoriamente de una muestra a otra y, por lo tanto, son a su vez variables aleatorias que tendrán su propia distribución de probabilidad que las describirá.

En los apartados que siguen vamos a estudiar la distribución de probabilidad de una **proporción** y de una **media**, cuando ambas son obtenidas a partir de los datos de una muestra. De la varianza obtenida en una muestra no nos ocuparemos por el momento.

2. Distribución de probabilidad de una proporción muestral

Son muchas y muy diversas las circunstancias en las que nos interesa conocer, o mejor estimar, la proporción de individuos en una población que comparten determinada característica: incidencia de determinada enfermedad, mujeres que comprarían determinado cosmético, jóvenes que consumen alcohol por encima de cierta cantidad, personas dispuestas a suscribir un nuevo tipo de póliza de seguros, …. Razones no nos faltan para estudiar cómo se comporta la proporción de ocurrencias de un suceso determinado en una muestra. Consideremos el siguiente ejemplo:

Ejemplo 3 Existe una teoría que asocia el desarrollo de actitudes agresivas con la presencia de una mutación del gen *bordus2-3*. Se sabe que dicha mutación está presente en el ADN del 10% de las personas. Si examinamos una muestra de 10 individuos elegidos al azar,

a) ¿qué valores puede tomar esa proporción?,
b) ¿con qué probabilidad tomará esos valores?,
c) ¿cúal es la probabilidad de que la proporción de individuos en los que está presente la mutación sea menor o igual que 0,4?

Observemos en primer lugar que a cada *proporción de individuos*, $\hat{\pi}$, poseyendo el gen mutante, le corresponde un determinado *número de individuos, Y*. En nuestro caso, con un tamaño de muestra igual a 10, esta correspondencia es la siguiente:

Y	0	1	2	3	4	5	6	7	8	9	10
$\hat{\pi}$	0,00	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	1,00

En consecuencia, es equivalente trabajar con una u otra característica.

Para contestar al apartado b), notemos que Y sigue una distribución binomial, B(10,0.1), porque se trata de una variable que

1. describe el número de ocurrencias (*éxitos*) de un suceso (*presencia del gen mutante*),
2. a lo largo de *n* pruebas independientes,
3. manteniéndose constante la probabilidad de ocurrencia en cada prueba, *p=0,1*.

Así, si queremos calcular las probabilidades que nos piden, escribiremos:

$$P(\hat{\pi} = 0) = P(Y = 0) = \binom{10}{0}(0,1)^0(0,9)^{10} = 0,3487$$

$$P(\hat{\pi} = 0,1) = P(Y = 1) = \binom{10}{1}(0,1)^1(0,9)^9 = 0,3874$$

$$P(\hat{\pi} = 0,2) = P(Y = 2) = \binom{10}{2}(0,1)^2(0,9)^8 = 0,1937$$

y así sucesivamente.

Finalmente, para obtener la probabilidad que nos piden en c), no tenemos más que recurrir a la tabla: $P(\hat{\pi} \le 0,4) = P(Y \le 4) = 0,9984$.

El razonamiento utilizado en el ejemplo nos permite generalizar el resultado:

Supongamos que extraemos una muestra aleatoria de tamaño *n* de una población en la que una proporción *p* de individuos posee determinada característica. Si el **número** y la **proporción** de individuos en la muestra con la mencionada característica los designamos mediante *Y* y $\hat{\pi}$, respectivamente, se verifica:

a) *Y* y $\hat{\pi}$ están ligados por la relación $\hat{\pi}$ = Y/n, y toman los valores

Y	0	1	2	……	n-2	n-1	n
$\hat{\pi}$	0/n	1/n	2/n	……	(n-2)/n	(n-1)/n	1

b) *Y* es una variable aleatoria **B(n,p)** y la distribución de probabilidad de $\hat{\pi}$ se obtiene mediante,

$$P(\hat{\pi} = \frac{k}{n}) = P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\cdots,n$$

Ejemplo 4 Se estima que el porcentaje de fumadores entre los adolescentes es del 40%. Estudiar la distribución de probabilidad de la proporción de fumadores que obtendremos al extraer, de la población de adolescentes, una muestra aleatoria de tamaño 6. Sabemos que el número de fumadores en la muestra será una variable aleatoria B(6,0.4) y a partir de la tabla de la Binomial, obtendremos las probabilidades asociadas a cada valor de $\hat{\pi}$.

Ejemplo 5 Un ejercicio interesante, y que nos permite interpretar mejor el significado de la distribución de probabilidad asociada a la variable $\hat{\pi}$, consiste en extraer muestras sucesivas todas del mismo tamaño, *n=6*. Hemos llevado a cabo la extracción de 200 de estas muestras y hemos descrito, en la tabla que sigue, la distribución de frecuencias de los valores de $\hat{\pi}$ obtenidos en todas estas muestras. Finalmente hemos comparado la frecuencia relativa con la probabilidad asociada a cada uno de estos valores, comprobando las escasas diferencias entre unos y otros, diferencias explicables por el error que todo proceso de muestreo comporta.

$\hat{\pi}$	Frec.	Frec. Rel. = f_k	$P(\hat{\pi}=k/n) = \hat{\pi}_k$	$\hat{\pi}_k - f_k$
0	12	0,060	0,0467	-0,0133
1/6	36	0,180	0,1866	0,0066
2/6	63	0,315	0,3110	-0,0040
3/6	50	0,250	0,2765	0,0265
4/6	25	0,125	0,1382	0,0132
5/6	11	0,055	0,0369	-0,0181
1	3	0,015	0,0041	-0,0109
Total	200	1,000	1,000	

Influencia del tamaño muestral Si *Y* es el número de individuos en la muestra que poseen la característica que nos interesa, *Y* se distribuye B(n,p). Recordemos que su media y su varianza valdrán, $\mu = np$ y $\sigma^2 = np(1-p)$. Como $\hat{\pi} = Y/n$, la media y varianza de $\hat{\pi}$ se obtiene fácilmente a partir de las anteriores,

$$\mu_{\hat{\pi}} = \frac{\mu_Y}{n} = p, \qquad \sigma^2_{\hat{\pi}} = \frac{\sigma^2_Y}{n^2} = \frac{p(1-p)}{n}.$$

De estos resultados deducimos dos interesantes propiedades:

1. La media de la proporción de individuos en la muestra que poseen la característica que nos interesa, $\hat{\pi}$, coincide con la proporción de los mismos en la población, *p*. Parece pues razonable **estimar** el verdadero valor, *p*, mediante el obtenido en la muestra, $\hat{\pi}$.
2. La varianza de $\hat{\pi}$, al ser inversamente proporcional al tamaño de la muestra, es tanto más pequeña cuanto mayor sea éste. Es decir, la dispersión de $\hat{\pi}$ alrededor de su media *p* disminuye cuando *n* aumenta. Dicho en otras palabras, la información que la muestra proporciona aumenta con su tamaño, lo que se corresponde, por otra parte, con lo que la intuición señala.

3. Distribución de probabilidad de una media muestral

La otra característica muestral que nos interesa estudiar es la media de los valores de la muestra, cuando la población de la que provienen es una variable aleatoria continua. Más concretamente, supondremos que la variable en estudio tiene una distribución de probabilidad **N (μ, σ^2)**. Es cierto que podemos estar interesados en extraer muestras de poblaciones correspondientes a otras distribuciones de probabilidad, pero el caso normal cubre la práctica totalidad de las situaciones que puedan presentársenos y satisface, por ello, el objetivo de este curso.

Extraemos una muestra de tamaño n de determinada característica, X, de una población. Si X tiene una distribución N (μ, σ^2), la media muestral, $\overline{X}$, es una variable aleatoria con distribución de probabilidad $N(\mu_{\overline{X}}, \sigma^2_{\overline{X}})$ cuyos parámetros valen, $\mu_{\overline{X}} = \mu$, $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.

Ejemplo 6 Si la variable X, que mide la puntuación global del test de delincuencia que mencionábamos al comienzo del tema, es N(75,144), ¿cuál es la probabilidad de que la media muestral, obtenida a partir de una muestra de tamaño 10, sea mayor que 80?
Sabemos que

$$\mu_{\overline{X}} = 75, \qquad \sigma^2_{\overline{X}} = \frac{144}{10}$$

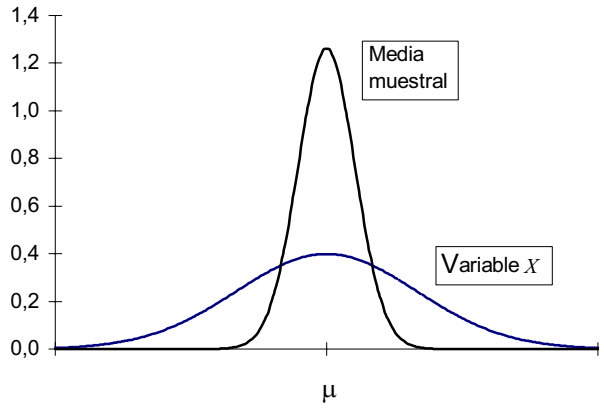
por tanto

$$P(\overline{X} > 80) = P\left(\frac{\overline{X} - 75}{\sqrt{144/10}} > \frac{80 - 75}{12/\sqrt{10}}\right) = P(Z > 1{,}32) = 1 - P(Z \le 1{,}32) = 0{,}0934 .$$

Lo que indica que, aproximadamente, en un 9.34% de las ocasiones la media muestral superará el valor 80.

Influencia del tamaño muestral También ahora la media y la varianza de la variable media muestral, $\overline{X}$, gozan de propiedades semejantes a las que tenía la variable proporción muestral. Recordemos que

$$\mu_{\overline{X}} = \mu, \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} .$$



De estos valores deducimos que:

1. La media muestral, cuya media coincide con la de la variable X de la cual hemos extraído la muestra, puede ser tomada como una **estimación** de la media de X.
2. La distribución de probabilidad de $\overline{X}$ tiene el mismo valor central que la de X, puesto que ambas comparten la misma media, μ, pero como su varianza es n veces menor que la de X, sus valores están mucho más concentrados. Concentración que aumenta a medida que lo hace el tamaño de la muestra o, dicho en otras palabras, la precisión de la información que la muestra proporciona aumenta con su tamaño. La gráfica adjunta nos ilustra esta propiedad.

TEMA 6.- ESTIMACIÓN Y CONTRASTE

1. Estimación de la media poblacional

Ya hemos señalado en el tema anterior que la imposibilidad de acceder a toda la población, con el fin de conocer el comportamiento aleatorio de una variable, puede paliarse accediendo a una parte de la población, **muestra**, y estudiando en ella las características que nos interesan. Los mecanismos para extender de la muestra a la población los resultados obtenidos nos los proporciona la **Inferencia Estadística** y decíamos que dos son los procesos fundamentales que llevaremos a cabo: **estimar** y **contrastar**. De ambos nos ocuparemos en este tema.

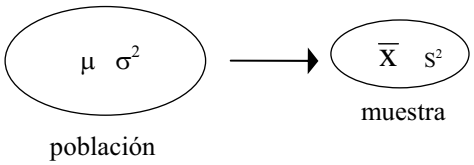
En el desarrollo del test sobre comportamiento delictivo, al que aludíamos en el tema anterior, los criminólogos han sometido al mismo a una muestra de 25 jóvenes, habiendo obtenido las siguientes puntuaciones:

76 82 72 63 76 80 79 64 73 83 78 45 86
66 81 82 97 55 73 70 85 73 75 78 86

a las que corresponden una media $\bar{\mathbf{x}} = 75{,}12$ y una varianza $\mathbf{s^2} = 128{,}46$. Investigaciones previas han permitido conocer que la variable aleatoria *X*, que describe el comportamiento del test en la población de jóvenes, tiene una distribución normal y conociendo los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\sigma^2}$ su distribución de probabilidad estará completamente determinada. Las observaciones de *X* obtenidas en la muestra permiten utilizar los **estadísticos muestrales** ($\bar{\mathbf{x}}$, $\mathbf{s^2}$) como valores aproximados de los **parámetros de la población** y decir que *X* es una variable aleatoria **N**(75.12,128.46). Es razonable llevar a cabo esta aproximación porque, recordemos, la media muestral es una variable aleatoria cuya media es precisamente μ. Para la varianza se verifica la misma propiedad y, en tanto que variable aleatoria, su media es σ^2. Esta manera de proceder podemos generalizarla.

La media y la varianza de una variable aleatoria *X* pueden **estimarse** a partir de los correspondientes valores obtenidos en una muestra:

$\bar{\mathbf{x}}$ es una **estimación** de μ,
$\mathbf{s^2}$ es una **estimación** de σ^2,
s es una **estimación** de σ.



2. Intervalo de confianza para la media de una población normal

A la estimación que hemos llevado a cabo en el párrafo anterior podemos objetar que, muy probablemente, una nueva muestra daría lugar a distintas media y varianza muestrales, lo que nos obligaría a admitir una distribución Normal para la variable *X* con distintos parámetros a los ya obtenidos. Una solución a la ambigüedad que esto supone consiste en acompañar la estimación de la media con alguna información adicional que recoja el carácter aleatorio de la media muestral (por lo que respecta a la varianza, el problema cae fuera del alcance de los objetivos de este curso). En efecto, si la variable original, *X*, se distribuye como una $N(\mu,\sigma^2)$, la media muestral tiene una distribución de probabilidad $N(\mu,\sigma^2/n)$ y podemos hacer uso de una interesante propiedad de la distribución normal, la que afirma que *la probabilidad de que el valor de una variable diste de su media a lo sumo dos veces su desviación típica es 0,95*. Aplicada a la variable $\bar{\mathbf{x}}$, con media μ y varianza σ^2/n:

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95\,.$$

Esto significa que en el 95% de las ocasiones el intervalo $I=\left[\,\mu \pm 2\frac{\sigma}{\sqrt{n}}\,\right]$ contendrá a $\bar{x}$.

Hemos mejorado nuestra información acerca de $\bar{x}$, pero nuestro objetivo principal sigue siendo μ. Observemos para ello que la distancia es una relación simétrica y que si $\bar{x}$ dista de μ a lo sumo $2\frac{\sigma}{\sqrt{n}}$, la distancia de μ a $\bar{x}$ será también menor o igual que dicha cantidad. Podemos por tanto escribir que

$$P\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

y disponer así de un intervalo que sabemos contiene a la media de la población con una probabilidad de 0,95.

No hemos resuelto nuestro problema definitivamente, porque para obtener los extremos del intervalo, $[\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}]$, necesitamos conocer el valor de σ y, en general, ésta es una característica de la población que suele ser desconocida. ¿Y si sustituimos, en las expresiones anteriores, σ por su estimación, *s*? La idea es lógica y aceptable, pero si lo hacemos y queremos que la probabilidad del 0,95 siga conservándose, hemos también de sustituir el factor 2 que multiplica a $\frac{\sigma}{\sqrt{n}}$ por una cantidad que denotaremos mediante $t_{1-0.95}$ y que está relacionada con una nueva distribución de probabilidad denominada **t de Student**.

Resumiendo, hemos obtenido un intervalo, $IC_{0.95}=[\bar{x} \pm t_{1-0.95}\frac{s}{\sqrt{n}}]$, al que denominaremos **intervalo de confianza al 95% para la media**, que tiene la siguiente propiedad:

$$P(\mu \text{ esté contenido en } IC_{0.95}) = 0.95$$

Observación Si hablamos de probabilidad es porque la ocurrencia o no del suceso $\{\mu$ *está contenido en* $IC_{0.95}\}$ depende del azar, pero esta aleatoriedad se deriva del carácter aleatorio de la longitud de $IC_{0.95}$, cuyos extremos vienen definidos en función de la variable aleatoria $\bar{x}$. Esta aleatoriedad del intervalo de confianza es fundamental para comprender su significado.

Error estándar de la media muestral Podemos observar que la longitud del intervalo depende de la variabilidad de la media muestral. Una medida de esta variabilidad es su desviación típica, $s/\sqrt{n}$, que recoge la dispersión de la media muestral respecto de la media poblacional debido al efecto de la muestra. Por esta razón se denomina también error muestral o **error estándar de la media muestral** a:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Ejemplo 1 Para comprender mejor el significado del intervalo de confianza y poner de manifiesto su aleatoriedad, supongamos que conocemos la distribución de probabilidad de la variable que mide la puntuación del test sobre delincuencia juvenil: N(75,144). Tomamos ahora 100 muestras de 10 jóvenes cada una, a los que sometemos al test y anotamos su puntuación. La tabla que sigue, que recoge los resultados para alguna de estas muestras, contiene los extremos superior e inferior del correspondiente intervalo de confianza al 95% y nos informa (columna **cubre?**) de si dicho intervalo contiene (**1**), o no (**0**), a la media poblacional, μ=75. Puesto que la probabilidad de que μ esté contenida en $IC_{0.95}$ es 0.95, cabe esperar, para un número de muestras suficientemente grande, que en el 95% de las ocasiones así ocurra. En nuestro caso, tal y como se indica al final de la tabla, 94 de los 100 intervalos, el 94%, contienen a μ=75.

				Intervalo de Confianza al 95%		
muestra	**media**	**desv. tip.**	$\mathbf{SE_{\bar{X}}}$	**ext. inf.**	**ext. sup.**	**cubre?**
1	71,96	9,30	2,94	65,30	78,61	1
2	78,11	14,58	4,61	67,69	88,54	1
3	72,13	14,70	4,65	61,61	82,65	1
4	70,50	12,82	4,05	61,33	79,66	1
……						
12	82,28	9,53	3,01	75,46	89,09	0
13	81,97	6,26	1,98	77,50	86,45	0
14	74,16	13,98	4,42	64,16	84,16	1
……						
97	72,76	11,83	3,74	64,29	81,22	1
98	75,28	8,55	2,70	69,16	81,39	1
99	74,43	14,93	4,72	63,75	85,11	1
100	65,17	7,83	2,48	59,57	70,77	0
					Total	**94**

Es lógico que construyamos los intervalos de confianza de tal forma que la probabilidad de que contengan a μ sea elevada, pero no tiene porqué coincidir siempre con 0,95. Podemos construir intervalos IC_q verificando

$$P(\mu \text{ esté contenido en } IC_q) = q,$$

para lo cual deberemos sustituir el anterior factor $t_{1-0.95}$ por un nuevo factor t_{1-q}, que ya dijimos que estaba ligado a la distribución de probabilidad conocida como **t de Student**. Esta distribución tiene la peculiaridad de que sus valores dependen de *n-1*, donde *n* es el tamaño de la muestra. En rigor, deberíamos haber escrito $t_{n-1,1-0.95}$ o $t_{n-1,1-q}$. Como viene siendo habitual, los valores de $t_{n-1,1-q}$ se pueden obtener de la correspondiente tabla, incluida como anexo al final del tema. En la tabla observamos que para valores de *n* muy grandes, señalados en la tabla como $n=\infty$, la distribución **t de Student** coincide con la **Normal**. Al valor *n-1* se le denomina **grados de libertad** de la distribución, de manera que nos referimos a ella como una **t de Student con *n-1* grados de libertad**.

Ejemplo 2 Para la muestra 1 de la tabla anterior, cuyas características son, *n=10*, $\bar{x}$ = 71.96 y s=9.30, los intervalos de confianza para distintos valores de *q* son:

q	$\mathbf{t_{9,1-q}}$	$\mathbf{SE_{\bar{X}}}$	$\mathbf{IC_q}$
0.90	1.833	2,94	[66.57, 77.34]
0.95	2.262	2,94	[65.30, 78.61]
0.99	3.250	2,94	[62.40, 81.51]

Se observa, como era lógico, que la longitud del intervalo aumenta con *q*. Si exigimos una mayor confianza, *q*, obtenemos una precisión menor en el intervalo que delimita a μ, mayor longitud, y viceversa.

3. Contraste de hipótesis para la media de una población Normal

Un ejemplo introductorio Unos laboratorios farmacéuticos desean adquirir una máquina de comprimir para fabricar comprimidos de 2 gramos de peso. El fabricante afirma que su máquina elabora comprimidos de peso medio 2 gramos, con una desviación típica de 0.025 gramos, siendo la distribución de probabilidad de la variable aleatoria *X*, peso del comprimido, Normal. Estas características, caso de ser ciertas, interesan a los laboratorios, que antes de tomar una decisión someten a la máquina a un control para comprobarlas. Para ello toman una muestra aleatoria de 25 comprimidos que da lugar a un peso medio de $\bar{x}$ = 1,99 gramos. ¿Qué decisión deberían tomar a la vista de este resultado?, o si se prefiere, ¿es este resultado compatible con la afirmación del fabricante?

Para tomar su decisión, los laboratorios podrían proceder de la siguiente forma: si la afirmación del fabricante es cierta, la media muestral, $\bar{x}$, tendrá una distribución de probabilidad $N(2,(0.025)^2/25)$. Como valores de la media muestral alejados de 2, por exceso o por defecto, pueden hacer poco creíble la afirmación del fabricante, y puesto que se ha obtenido $\bar{x}$ =1,99, es conveniente calcular la probabilidad de obtener valores semejantes. En concreto, calcularemos la probabilidad de obtener valores cuya media muestral sea a lo sumo la observada: $P(\bar{x} \le 1,99)$,

$$P(\bar{x} \le 1,99) = P\left(\frac{\bar{x} - 2}{0.025/\sqrt{25}} \le \frac{1,99 - 2}{0,005}\right) = P(Z \le -2) = 0,0228 \approx 0,023$$

Probabilidad muy baja, cuya interpretación supone que, aproximadamente, sólo el 2,30% de las muestras que obtengamos conducirán a medias muestrales menores o iguales que 1,99, siempre bajo la hipótesis de creer al fabricante.

Una probabilidad tan baja puede conducirnos a esta doble reflexión:

1. Es cierto que nos ha ocurrido un suceso, $\{\bar{x} \le 1,99\}$, que tiene muy baja probabilidad, pero quizás haya dado *la casualidad* de que hayamos elegido para la muestra alguno de los *muy poco probables* resultados que no son favorables a lo que afirma el fabricante. Este sería un planteamiento *optimista*, porque supone creer que nos ha ocurrido algo muy improbable, que conduce a aceptar lo que dice el fabricante y, en consecuencia, a adquirir su máquina.

2. Un razonamiento menos optimista, pero sin duda más *realista*, supone admitir la posibilidad de que el suceso ha ocurrido porque su probabilidad es, en realidad, mayor que la obtenida bajo la hipótesis del fabricante, cosa que sólo es posible si rechazamos esa hipótesis. En efecto, si a la vista del resultado obtenido supusiéramos que el peso medio de los comprimidos es, por ejemplo, 1,995 gramos, tendríamos que $P(\bar{x} \leq 1{,}99)=P(Z \leq -1)=0.1587$, que haría más comprensible el resultado muestral obtenido.

Una decisión prudente sería rechazar lo que el fabricante afirmó y no adquirir la máquina que nos ofrecen, porque el peso medio de los comprimidos que fabrica parece ser distinto de los 2 gramos anunciados por su fabricante.

La situación anterior constituye un ejemplo de lo que en Inferencia Estadística denominamos **contraste de hipótesis**. El ejemplo contiene todos los elementos característicos de un contraste y describe el procedimiento a seguir. Extraigamos los rasgos esenciales del procedimiento de contraste.

Condiciones iniciales Las condiciones iniciales suponen que la variable aleatoria *X*, objeto de estudio, sigue una distribución $N(\mu,\sigma^2)$, con σ^2 desconocida.

Hipótesis nula e hipótesis alternativa En cualquier problema de contraste de hipótesis debemos comenzar por establecer claramente cuál es la hipótesis acerca de μ que queremos contrastar, a la que se denomina **hipótesis nula** y se la designa mediante $\mathbf{H_0}$. Deberemos también establecer una **hipótesis alternativa**, $\mathbf{H_A}$, que será la finalmente aceptada si los datos de la muestra conducen al rechazo de H_0. El siguiente esquema nos muestra el establecimiento de H_0 y H_A en el ejemplo anterior y en general:

Ejemplo	**En general**
$\mathbf{H_0}$: $\mu = 2$	$\mathbf{H_0}$: $\mu = \mu_0$
$\mathbf{H_A}$: $\mu \neq 2$	$\mathbf{H_A}$: $\mu \neq \mu_0$

El estadístico t_s y su P-valor asociado El valor a partir del cual vamos a tomar la decisión de aceptar H_0 o su alternativa, H_A, es la media muestral. Pero la simple observación de su valor no es suficiente para decidir. La decisión, como hemos visto en el ejemplo, la tomamos una vez conocemos la probabilidad de obtener valores tan extremos como el que la media muestral nos presenta. Esta probabilidad puede calcularse *si suponemos que H_0 es cierta*, pues entonces la media muestral, $\bar{x}$, tiene una distribución $N(\mu_0,\sigma^2/n)$. Así, si el valor concreto que toma la media muestral es $\bar{x}=\bar{x}_0$, la probabilidad que nos interesa se obtiene fácilmente a partir del valor tipificado,

$$z_0 = \frac{\bar{x}_0 - \mu_0}{\sigma/\sqrt{n}}.$$

Ocurre, al igual que en la obtención de los intervalos de confianza, que no conocemos el verdadero valor de σ y al sustituirlo en la expresión anterior por su estimación, la desviación típica muestral *s*, lo que obtenemos es un valor, al que denominamos **estadístico t_s**, que se distribuye como una t de Student con *n-1* grados de libertad:

$$\boxed{t_s = \frac{\bar{x}_0 - \mu_0}{s/\sqrt{n}}}$$

Asociado al **estadístico t_s**, existe lo que denominamos **P-valor**, que es la probabilidad de obtener valores del estadístico tan extremos como t_s, es decir

$$\text{P-valor} = 2\cdot P(T \geq |t_s|),$$

donde T es una variable aleatoria t de Student con *n-1* grados de libertad. Usamos $|t_s|$ porque al establecer H_A como $\mu \neq \mu_0$, admitimos la posibilidad de que la muestra nos proporcione valores mayores o menores que μ_0 y por tanto t_s puede ser positivo o negativo. Para obtener el P-valor recurrimos a la tabla de la t de Student buscando $|t_s|$ en la entrada correspondiente a *n-1*, de no

encontrarlo recurrimos a los dos valores más próximos por defecto y por exceso. El P-valor será el valor que encabeza la correspondiente columna, o estará comprendido entre los dos valores que encabezan las columnas correspondientes a los valores más próximos por defecto y por exceso.

Nivel de significación Nuestra decisión final depende del P-valor. Decidiremos de acuerdo con la siguiente regla:

- rechazamos $\mathbf{H_0}$, lo que equivale a aceptar $\mathbf{H_A}$, si el P-valor es pequeño,
- no rechazamos $\mathbf{H_0}$ si el P-valor es grande.

Pero así planteada es una regla muy imprecisa y, por tanto, de escasa utilidad en la práctica. Hemos de fijar un valor que establezca claramente la división entre P-valor grande y pequeño; a dicho valor se le denomina **nivel de significación** y se le designa mediante la letra griega α. Habitualmente se asigna a α el valor 0.05, razón por la cual en el ejemplo anterior, con un P-valor = 2x0.023 = 0.046, hemos rechazado H_0. En ocasiones se utilizan otros valores para α, por ejemplo, $\alpha = 0.10$ o $\alpha = 0.01$, dependiendo ello del problema en estudio. La elección del nivel de significación se lleva a cabo al comienzo del proceso de contraste, por ejemplo cuando establecemos las hipótesis nula y alternativa.
Fijado el nivel de significación, la regla de decisión anterior queda completamente fijada:

Regla de decisión en función del P-valor	
si **P-valor** > α	no rechazamos H_0
si **P-valor** ≤ α	rechazamos H_0

Un último comentario. No olvidemos que hemos tomado nuestra decisión a partir de la información que nos proporciona la muestra y que el error que todo proceso de muestreo comporta puede hacer posible que hayamos obtenido un valor extremo, aparentemente incompatible con una H_0 que, sin embargo, es cierta. Esto permite dar al nivel de significación otra interesante interpretación: *el nivel de significación es el error que cometemos cuando rechazamos una H_0 que es cierta*. Esta es una nueva razón para que α sea pequeño.

Una vez descritos los distintos elementos que intervienen en el proceso de contrastar la media de una población Nornal es conveniente resumir y ordenar las distintas acciones que hemos de llevar a cabo.

RESUMEN: CONTRASTE PARA LA MEDIA DE UNA POBLACION

1. La variable aleatoria X tiene una distribución $N(\mu,\sigma^2)$, con σ^2 desconocida
2. Establecemos las hipótesis acerca μ:
   $\mathbf{H_0}$: $\mu = \mu_0$
   $\mathbf{H_A}$: $\mu \neq \mu_0$
3. Elegir el nivel de significación, habitualmente $\boldsymbol{\alpha = 0.05}$
4. Elegir una muestra de tamaño **n** y calcular $\bar{x}$ y $\mathbf{s^2}$
5. Obtener el **estadístico $\mathbf{t_s}$** mediante la fórmula
$$t_s = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
6. Comparar el **P-valor** correspondiente a $\mathbf{t_s}$ y el **nivel de significación**, α, y decidir lo que corresponda de acuerdo con la regla de decisión ya conocida:
   si **P-valor** > α aceptamos H_0
   si **P-valor** ≤ α rechazamos H_0

Ejemplo 4 Cuando el test sobre comportamiento delictivo se aplica a la población juvenil en general, la puntuación, X, del test se comporta como una Normal de media $\mu = 75$. Existen razones para pensar que las puntuaciones obtenidas en el test por los jóvenes con mayor propensión a la delincuencia no se comportan de la misma manera. Para comprobar esta teoría hemos tomado una muestra de 30 de estos jóvenes, cuyas puntuaciones han sido:

62 86 85 92 75 80 78 78 88 73 73 96 72 97 76
97 90 101 66 72 73 86 84 66 86 74 69 65 81 70

¿Qué podemos decir a la vista de estos resultados?

Plantearemos el contraste siguiendo los pasos señalados en el resumen anterior:

1. Por lo que respecta a las condiciones iniciales, el problema nos dice que la variable aleatoria *X*, puntuación obtenida en el test, es $N(75,\sigma^2)$. Observemos que, como nada se afirma, la varianza es desconocida.
2. Las hipótesis que se derivan del problema planteado son:

   H_0: $\mu = 75$

   H_A: $\mu \neq 75$
3. Elegiremos el **nivel de significación** habitual $\alpha = 0.05$.
4. De la muestra de las puntuaciones obtenidas por los 30 jóvenes calculamos

$$\bar{x}_0 = 79.7 \quad s^2 = 112.15 \quad s = 10.59$$

5. El valor del **estadístico t_s** asociado a $\bar{x}_0$ y s^2 es

$$t_s = \frac{79.7 - 75}{10.59/\sqrt{30}} = 2.43$$

6. Al entrar en la tabla de la t de Student por la fila correspondiente a los 29 grados de libertad, comprobamos que el **P-valor** asociado a $|t_s| = 2.43$ verifica

0.05	**P-valor**	**0.02**
2.045	2.43	2.462

   por lo que **0.05 > P-valor > 0.02** y, de acuerdo con la regla de decisión establecida, rechazaremos H_0 y aceptaremos H_A. Concluimos que los datos que la muestra nos proporciona no están en contradicción con la teoría que postula un comportamiento distinto de la puntuación en el test por parte de los jóvenes con propensión a la delincuencia.

Otro tipo de hipótesis alternativas A la vista del resultado de la muestra hubiera sido más lógico plantear como hipótesis alternativa

$$H_A: \quad \mu > 75.$$

Esta es una situación que se plantea con frecuencia en muchos problemas de contraste, que requieren una hipótesis alternativa que señale la dirección en la que la desigualdad se produce. Dos son las posibilidades que tenemos de plantear hipótesis alternativas **direccionales**, que así son conocidas,

$$H_A: \ \mu > \mu_0 \qquad \text{y} \qquad H_A: \ \mu < \mu_0$$

En estas situaciones, hemos de modificar el procedimiento anterior a través de los dos pasos siguientes:

- **Paso 1** Comprobamos la direccionalidad viendo si la media muestral se desvía de μ_0 en la dirección señalada por H_A
  a) en caso negativo, el P-valor > 0.50 y aceptamos H_0 sin necesidad de continuar
  b) en caso afirmativo, llevamos a cabo el segundo paso
- **Paso 2** Continuamos el proceso tal y como lo hemos descrito para el caso **no direccional** (H_A: $\mu \neq \mu_0$), hasta calcular el estadístico t_s. La diferencia estriba en que dividiremos por 2 el P-valor que obtengamos, siendo el valor resultante el que utilizaremos como P-valor en la regla de decisión.

  Ejemplo 5 Retomemos el ejemplo 4 pero planteemos ahora las hipótesis de esta forma:

  H_0: $\mu = 75$

  H_A: $\mu > 75$

  - **Paso 1** Como la media obtenida en la muestra es $\bar{x}_0 = 79.7 > 75$, confirma la dirección señalada por H_A, y debemos continuar
  - **Paso 2** Procedemos como lo hicimos en el ejemplo 4: obtenemos el valor del estadístico t_s=2.43, pero el P-valor asociado verifica ahora

0.025	P-valor	0.01
2.045	2.43	2.462

y, por tanto, **0.025 > P-valor > 0.01**, aceptando la hipótesis alternativa de una mayor puntuación media en los test correspondientes a jóvenes con propensión a la delincuencia.

Si la media muestral hubiera sido 74, no hubiera corroborado la dirección de la hipótesis alternativa, haciéndose innecesario seguir porque semejante valor nunca va a darnos evidencia a favor de la hipótesis alternativa.

4. Comparación de medias de dos poblaciones Normales

Un ejemplo introductorio El nivel de hematocrito es una medida de la concentración de glóbulos rojos en sangre. En la tabla se muestran los valores del hematocrito correspondientes a dos muestras de jóvenes de 18 años, 20 hombres y 22 mujeres. Las muestras han sido tomadas independientemente una de otra.

hombres		mujeres	
46,62	39,87	36,86	41,66
46,67	39,87	38,17	42,00
45,01	46,21	48,12	43,89
44,68	44,11	44,42	43,90
47,75	46,67	43,08	39,49
36,63	41,94	38,79	47,71
42,17	42,52	35,87	40,32
44,13	47,53	37,24	44,15
40,78	48,50	42,91	40,82
40,79	42,52	43,65	39,73
		42,05	35,07

Sabemos que las variables aleatorias que miden el hematocrito en las poblaciones de mujeres y hombres tienen una distribución de probabilidad Normal con parámetros,

$$X_M \sim N(\mu_1, \sigma_1^2) \qquad X_H \sim N(\mu_2, \sigma_2^2),$$

cuya estimación, de acuerdo con lo que dijimos anteriormente, podemos obtenerla a partir de las correspondientes medias y varianzas muestrales:

mujeres: $\overline{x}_1 = 41{,}36 \quad s_1^2 = 12{,}31$

hombres: $\overline{x}_2 = 43{,}75 \quad s_2^2 = 10{,}21$

Podemos también construir los IC_q para las medias de cada población. Pero aun teniendo interés toda esta información, datos de esta naturaleza conducen de inmediato a preguntarse, *¿tienen hombres y mujeres el mismo nivel medio de hematocrito o, tal como las muestras parecen sugerir, es distinto de unos a otros?* Lo que nos lleva al problema de **comparar las medias de dos poblaciones Normales**.
Observemos que comparar medias equivale a contrastar alguna hipótesis acerca de su diferencia. En efecto, en el ejemplo anterior la hipótesis $\mu_1 = \mu_2$ es equivalente a $\mu_1 - \mu_2 = 0$. Nuestro problema se reducirá por tanto a estimar, construir intervalos de confianza y contrastar la diferencia de ambas medias. Habrá que distinguir dos situaciones: que las **varianzas** de las poblaciones sean **iguales** o **distintas**.

Caso de igualdad de varianzas poblacionales: $\sigma_1^2 = \sigma_2^2$ Sigamos el esquema establecido para el caso de una población.

Condiciones iniciales Recordemos una vez más los supuestos iniciales en los que se basan los procedimientos que vamos a desarrollar a continuación. Estamos en presencia de dos poblaciones, es decir, de dos variables aleatorias, ambas Normales e independientes:

$$X_1 \sim N(\mu_1, \sigma_1^2) \qquad X_2 \sim N(\mu_2, \sigma_2^2),$$

con $\sigma_1^2 = \sigma_2^2$. Hemos extraído sendas muestras independientes, de tamaños, n_1 y n_2, respectivamente.

Estimación de $\mu_1 - \mu_2$ Puesto que cada una de las medias poblaciones se estiman mediante sus correspondientes medias muestrales,

$\bar{x}_1 - \bar{x}_2$ es una estimación de $\mu_1 - \mu_2$

Intervalo de confianza para $\mu_1 - \mu_2$ A semejanza de los intervalos que hemos construido para la media de una población, el que construyamos para la diferencia $\mu_1 - \mu_2$ tendrá la forma

$$IC_q = [\,(\bar{x}_1 - \bar{x}_2) \pm t_{GL,1-q} \cdot SE_{(\bar{x}_1 - \bar{x}_2)}\,]$$

y su punto medio será la diferencia de medias muestrales. Hay, lógicamente, algunas diferencias respecto de la situación anterior, las que se derivan de trabajar ahora con dos poblaciones en lugar de con una. La primera reside en el número de grados de libertad de la t de Student que ahora se obtiene a partir del tamaño de ambas muestras:

$$GL = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2.$$

Otra diferencia estriba en la obtención del valor del error estándar de $\bar{x}_1 - \bar{x}_2$, cuyo valor es ahora

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{s_T^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

siendo

$$s_T^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

La expresión final de IC_q será:

$$IC_q = [\,(\bar{x}_1 - \bar{x}_2) \pm t_{GL,1-q} \cdot \sqrt{s_T^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

verificándose, $P(\mu_1 - \mu_2$ esté contenido en $IC_q) = q$.

Contraste de hipótesis para la igualdad de medias de dos poblaciones La estructura del contraste es la misma que la desarrollada para el caso de una población. En consecuencia, podemos pasar directamente al resumen final, adecuadamente adaptado:

RESUMEN: COMPARACIÓN DE DOS MEDIAS (Varianzas iguales)

1. Las dos variables aleatorias son Normales e independientes

   $X_1 \sim N(\mu_1, \sigma_1^2) \qquad X_2 \sim N(\mu_2, \sigma_2^2),$

   con $\sigma_1^2 = \sigma_2^2$.

2. Las hipótesis que se establecen acerca de $\mu_1 - \mu_2$ son:
   H_0: $\mu_1 - \mu_2 = 0$, o su equivalente $\mu_1 = \mu_2$
   H_A: $\mu_1 \neq \mu_2$
3. Elegir el nivel de significación, habitualmente **$\alpha = 0.05$**

4. Elegidas las muestras de tamaño $\mathbf{n_1}$ y $\mathbf{n_2}$, respectivamente, calcular $\bar{\mathbf{x}}_\mathbf{1}$, $\bar{\mathbf{x}}_\mathbf{2}$, $\mathbf{s_1^2}$ y $\mathbf{s_2^2}$

5. Obtener el correspondiente **estadístico t_s** mediante la fórmula

$$t_s = \frac{\bar{x}_1 - \bar{x}_2}{SE_{(\bar{x}_1 - \bar{x}_2)}},$$

donde

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{s_T^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \text{ y } s_T^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)},$$

recordando que los grados de libertad son ahora $\mathbf{GL = n_1 + n_2 - 2}$

6. Comparar el **P-valor** asociado a $|t_s|$ y **el nivel de significación**, α, y decidir lo que corresponda de acuerdo con la regla de decisión:

   si **P-valor** $> \alpha$ aceptamos H_0

   si **P-valor** $\leq \alpha$ rechazamos H_0

Ejemplo 6 Estamos ahora en condiciones de efectuar el contraste que proponíamos al principio del párrafo. Los valores de las varianzas de las observaciones y la naturaleza el problema permiten suponer que las varianzas de ambas poblaciones serán iguales.

1. Las hipótesis a establecer son:

   H_0: $\mu_1 = \mu_2$

   H_A: $\mu_1 \neq \mu_2$

2. Elegiremos el **nivel de significación** habitual $\alpha = 0.05$

3. De las dos muestras de hematocritos en hombres y mujeres de 18 años calculamos

   mujeres: $\bar{x}_1 = 41{,}36$ $s_1^2 = 12{,}31$

   hombres: $\bar{x}_2 = 43{,}75$ $s_2^2 = 10{,}21$

4. Para obtener el valor del **estadístico t** asociado a estos valores,

$$s_T^2 = \frac{21 \times 12{,}31 + 19 \times 10{,}21}{40} = 11{,}31$$

   y

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{11{,}31 \times \left(\frac{1}{22} + \frac{1}{20}\right)} = \sqrt{1{,}0796} = 1{,}04,$$

   y finalmente,

$$t_s = \frac{41{,}36 - 43{,}75}{1{,}04} = -2{,}30,$$

   y los grados de libertad de la correspondiente t de Student, $\mathbf{GL = 20 + 22 - 2 = 40}$.

6. Al entrar en la tabla de la t de Student por la fila correspondiente a los 40 grados de libertad, comprobamos que el **P-valor** asociado a $|t_s| = 2.30$ verifica

0.05	**P-valor**	**0.02**
2.021	2.30	2.457

   por lo que **0.05 > P-valor > 0.02** y, de acuerdo con la regla de decisión establecida, rechazaremos H_0 y aceptaremos la hipótesis alternativa, H_A, de un nivel medio de hematocrito distinto en hombres y mujeres.

Caso de varianzas poblacionales distintas $\sigma_1^2 \neq \sigma_2^2$ La única diferencia con el caso de igualdad de varianzas reside en el valor del error estándar de $\bar{x}_1 - \bar{x}_2$. Bajo el supuesto de

varianzas poblacionales distintas, el cuadrado del error estándar de la diferencia de medias muestrales es la suma de los cuadrados de los errores estándar de cada una de ellas,

$$SE^2_{(\bar{x}_1-\bar{x}_2)} = SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}.$$

Es decir,

$$SE_{(\bar{x}_1-\bar{x}_2)} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}}.$$

La expresión final de IC_q será:

$$IC_q=\left[(\bar{x}_1 - \bar{x}_2) \pm t_{GL,1-q}\cdot \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} \right],$$

verificándose, $P(\mu_1 - \mu_2$ esté contenido en $IC_q) = q$.

Contraste de hipótesis para la igualdad de medias de dos poblaciones La estructura del contraste es la desarrollada anteriormente. Su resumen, adecuadamente adaptado, es el siguiente:

RESUMEN: COMPARACIÓN DE DOS MEDIAS (Varianzas distintas)

1. Las dos variables aleatorias son Normales e independientes

   $X_1 \sim N(\mu_1, \sigma_1^2)$ $\quad X_2 \sim N(\mu_2, \sigma_2^2)$,

   con $\sigma_1^2 \neq \sigma_2^2$.

2. Las hipótesis que se establecen acerca de $\mu_1 - \mu_2$ son:
   H_0: $\mu_1 - \mu_2 = 0$, o su equivalente $\mu_1 = \mu_2$
   H_A: $\mu_1 \neq \mu_2$
3. Elegir el nivel de significación, habitualmente $\alpha = 0.05$
4. Elegidas las muestras de tamaño n_1 y n_2, respectivamente, calcular $\bar{x}_1$, $\bar{x}_2$, s_1^2 y s_2^2
5. Obtener el correspondiente **estadístico t_s** mediante la fórmula

   $$t_s = \frac{\bar{x}_1 - \bar{x}_2}{SE_{(\bar{x}_1-\bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

   recordando que los grados de libertad son ahora $GL = n_1 + n_2 - 2$
6. Comparar el **P-valor** asociado a $|t_s|$ y **el nivel de significación**, α, y decidir lo que corresponda de acuerdo con la regla de decisión:
   si **P-valor** $> \alpha$ aceptamos H_0
   si **P-valor** $\leq \alpha$ rechazamos H_0

Ejemplo 7 Se quiere estudiar el efecto de una misma dieta alimenticia sobre animales distintos pero de características somatométricas similares. Se eligen para ello ratas y hámsteres y observando el incremento de peso de cada animal entre los días 30 y 60 del proceso. Los resultados se recogen en la tabla adjunta.

Ratas		**Hámsteres**	
134	161	70	100
146	107	101	85
104	83	107	105
119	113	94	
124	129		
97	123		

En un problema de estas características lo razonable es suponer la desigualdad de varianzas poblacionales.

1. Las hipótesis a establecer son:

$$\mathbf{H_0:} \quad \mu_1 = \mu_2$$

$$\mathbf{H_A:} \quad \mu_1 \neq \mu_2$$

2. Elegiremos el **nivel de significación** habitual $\alpha = 0.05$
3. De las dos muestras calculamos

$$\textbf{hámsteres:} \quad \bar{x}_1 = 94{,}57 \quad s_1^2 = 171{,}62$$

$$\textbf{ratas:} \quad \bar{x}_2 = 120{,}00 \quad s_2^2 = 457{,}45$$

4. Para obtener el valor del **estadístico t** asociado a estos valores,

$$t_s == \frac{94{,}57 - 120}{\sqrt{\frac{171{,}62}{7} + \frac{457{,}45}{12}}} = -3{,}21,$$

y los grados de libertad de la correspondiente t de Student, **GL = 7 + 12 -2 = 17**.

6. Al entrar en la tabla de la t de Student por la fila correspondiente a los 17 grados de libertad, comprobamos que el **P-valor** asociado a $|\mathbf{t_s}| = 3{,}21$verifica

0.01	**P-valor**	**0.001**
2.898	3.21	3.965

por lo que **0.01 > P-valor > 0.001** y, de acuerdo con la regla de decisión establecida, rechazaremos H_0 y aceptaremos la hipótesis alternativa, H_A, de un incremento de peso distinto en ratas y hámsteres sometidos a la misma dieta.

Nota ¿Cómo decidir si usamos el procedimiento correspondiente a $\sigma_1^2 = \sigma_2^2$ o $\sigma_1^2 \neq \sigma_2^2$? Como σ_1^2 y σ_2^2 son generalmente desconocidos, hemos de basarnos en sus estimadores s_1^2 y s_2^2 . Una regla que se suele utilizar en la práctica es considerar que las varianzas son iguales (y por tanto utilizaremos el primer procedimiento) si la diferencia entre s_1 y s_2 no excede el 30% del menor de ambos. En caso contrario, se considera que son distintas y se aplica el segundo procedimiento. Fijémonos que en el ejemplo 7

$$s_1^2 = 171{,}62, \text{ con lo que } s_1 = 13.1, \text{ y}$$

$$s_2^2 = 457{,}45, \text{ con lo que } s_2 = 21.4,$$

su diferencia vale 21.4 – 13.1 = 8.3, mayor que 3.93 que corresponde al 30% de 13.1. Razón por la cual el ejemplo se resolvió aplicando el segundo procedimiento.

Otro tipo de hipótesis alternativas Al igual que en el caso de una población, pueden surgir situaciones que aconsejen plantear hipótesis alternativas direccionales. En el ejemplo del hematocrito, a la vista del resultado de la muestra, hubiera sido más conveniente plantear como alternativa a la igualdad de niveles medios en hombres y mujeres, esta otra hipótesis

$$\mathbf{H_A:} \quad \mu_1 < \mu_2$$

Las dos posibles alternativas **direccionales** son ahora:

$$\mathbf{H_A:} \quad \mu_1 < \mu_2$$

$$\mathbf{H_A:} \quad \mu_1 > \mu_2$$

y el procedimiento se modifica de forma análoga a como hicimos para el caso de una población.

Ejemplo 8 Retomemos el ejemplo 6 planteando ahora las hipótesis de esta forma:

$$\mathbf{H_0:} \quad \mu_1 = \mu_2$$

$$\mathbf{H_A:} \quad \mu_1 < \mu_2$$

- **Paso 1** Como las medias obtenidas en las muestras confirman la dirección señalada por H_A, puesto que $\overline{x}_1 < \overline{x}_2$ ($\overline{x}_1 = 41{,}36$ y $\overline{x}_2 = 43{,}75$), debemos continuar

- **Paso 2** Procedemos como lo hicimos en el ejemplo 6, obtenemos el valor del estadístico $|\mathbf{t_s}|=2{,}30$, pero el P-valor asociado verifica ahora

0.025	**P-valor**	**0.01**
2.021	2.30	2.457

y, por tanto, **0.025 > P-valor > 0.01**, lo que nos induce a rechazar H_0 y aceptar que el nivel medio de hematocrito es mayor en los hombres.