CONOCIMIENTO EDUCATIVO

Instituto Superior de Formación del Profesorado

La Estadística y la Probabilidad en el Bachillerato

MINISTERIO
DE EDUCACIÓN,
CULTURA Y DEPORTE

LA ESTADÍSTICA Y LA PROBABILIDAD EN EL BACHILLERATO

MINISTERIO
DE EDUCACIÓN,
CULTURA Y DEPORTE

SECRETARÍA GENERAL
DE EDUCACIÓN Y
FORMACIÓN PROFESIONAL

INSTITUTO SUPERIOR
DE FORMACIÓN
DEL PROFESORADO

MINISTERIO DE EDUCACIÓN, CULTURA Y DEPORTE
SECRETARÍA GENERAL DE EDUCACIÓN Y FORMACIÓN PROFESIONAL
Instituto Superior de Formación del Profesorado

Edita:

Subdirección General de Información y Publicaciones

N.I.P.O.: 176-03-005-0
I.S.B.N.: 84-369-3660-4
Depósito legal: M. 4.597-2003

Imprime: Fareso, S. A.

Cualquier docente que disponga de una conexión a Internet y de un buscador habrá podido comprobar que existe una gran cantidad de material en la red relacionado con los temas que se abordan en este libro, muchos de ellos incluyendo animaciones que dinamizan el desarrollo de las unidades que describen.

Cuando se elaboraron estos materiales, hace casi siete años, los medios que permiten realizar presentaciones y ejemplificaciones dinámicas estaban muy poco difundidos en los ambientes educativos, no por desconocimiento sino, en buena parte de los casos, por insuficiencia de recursos. Qué duda cabe que, si hoy día hubiéramos tenido que producir estas unidades, el enfoque didáctico hubiera sido diferente.

Con todo, el tiempo no ha producido, que sepamos, nuevos enfoques sobre la forma de entender los contenidos que aquí se presentan, ni nuevas aportaciones que obliguen a cambiar el trazado que discurre desde la organización estadística de los datos hasta las cuestiones inferenciales que ayudan a asignar a dichos datos un modelo estocástico de comportamiento.

Por ello, creemos que el trabajo puede seguir resultando interesante, tanto por las propuestas que formula sobre unidades didácticas (aunque, obviamente, deben ser complementadas con material dinámico) como por las sugerencias que propone al profesor acerca de la mirada con que debe acercarse a las cuestiones inferenciales.

Si el lector, especialmente el docente, participa de esta opinión, nos sentiremos acompañados en nuestro trabajo. También nos sentiremos acompañados si es crítico con estos materiales y decide comunicarnos su enfoque y sus críticas.

El conjunto de la obra consta de dos volúmenes. Este segundo se dedica a los materiales que, en nuestra opinión, formarían parte del currículum de los distintos bachilleratos. Como se verá, a diferencia de las propuestas para la E.S.O., aquí no se propone una secuenciación temporal de los contenidos, ya que la diversidad de

programaciones docentes haría la descripción muy farragosa. En su lugar, se busca una coherencia interna del discurso, al menos en lo referente al empleo de las distribuciones de probabilidad y a la Inferencia estadística.

La Introducción, cuya lectura consideramos muy conveniente, ayudará a valorar el alcance y los objetivos de los materiales, cuyo nivel intenta ser el que se busca para el alumno excepto en el capítulo correspondiente a los Test de contraste de hipótesis, en buena parte orientado hacia la formación del profesorado.

Uno de los miembros del equipo, Eduardo Fernández Amatria, nos ha abandonado en la primavera de 2002, cuando el material tomaba su forma actual. A él, siempre crítico y siempre colaborador, se lo dedicamos de una forma muy especial.

Colección: CONOCIMIENTO EDUCATIVO
Serie: Didáctica

LA ESTADÍSTICA Y LA PROBABILIDAD EN EL BACHILLERATO

Estos materiales y los del volumen anterior (*La Estadística y la Probabilidad en la Educación Secundaria Obligatoria*) han sido elaborados dentro del proyecto de cooperación entre departamentos universitarios y de enseñanza secundaria LA ESTADÍSTICA Y LA PROBABILIDAD EN LOS CURRÍCULOS DE LA EDUCACIÓN SECUNDARIA OBLIGATORIA Y EL BACHILLERATO, que ha sido promovido por el Departamento de Economía Aplicada (Estadística y Econometría) de la Universidad de Valladolid y los Departamentos de Matemáticas de los Institutos de Educación Secundaria Emilio Ferrari, Vega del Prado y Leopoldo Cano, los tres de Valladolid.

El proyecto ha sido realizado con una subvención del Ministerio de Educación y Cultura (concurso convocado por la Secretaría de Estado de Educación por Resolución de 15 de febrero de 1995, B.O.E. del 24 de febrero, y resuelto por Resolución de 24 de julio de 1995, B.O.E. de 11 de septiembre, de la citada Secretaría de Estado) finalizando su ejecución en septiembre de 1996.

Dirección editorial del volumen *La Estadística y la Probabilidad en el Bachillerato*: JOSÉ LUIS ROJO GARCÍA

Coordinación: *MUÑOZ VEGA, José Antonio.*

Autores:

BARBERO SANPEDRO, Carmen.
BERBEL HERNÁNDEZ, María.
FERNÁNDEZ AMATRIA, Eduardo.
FERNÁNDEZ DE LA MORA, Julia.
FERNÁNDEZ-ABASCAL TEIRA, Hermenegildo.
MARTÍN ROJO, Isabel.
MATEO AYUSO, María Luz.
MUÑOZ VEGA, José Antonio.
NEGUERUELA SÁNCHEZ, Isabel.
REDONDO GONZÁLEZ, Rosa.
ROJO GARCÍA, José Luis.
SANZ GÓMEZ, José Antonio.
TAMAMES RODRÍGUEZ, Luis.
VÁZQUEZ RODRÍGUEZ, Reinerio.
ZARZOSA ESPINA, Pilar.

ÍNDICE

INTRODUCCIÓN

Se presentan a continuación los materiales referentes a los distintos bachilleratos. Abordan cuestiones relacionadas con la *Estadística descriptiva,* y en concreto, el análisis bidimensional, con especial referencia a la regresión lineal.

Más adelante, realizamos una introducción de los conceptos de *probabilidad.* Como puede verse, el estudio se limita a los espacios muestrales discretos, ya que en ellos no se plantean problemas relacionados con la selección de sucesos y, para los objetivos que se pretenden, resultan suficientes.

A continuación se definen las variables aleatorias, y las herramientas que permiten manejarlas sin dificultad. El cálculo de probabilidades para variables continuas se plantea de un modo puramente instrumental, de forma que el alumno sea capaz de seguir los cálculos relacionados con variables normales.

Dentro de las cuestiones relacionadas con la *Inferencia estadística,* se abordan los problemas de la estimación puntual y, relacionada con ellos, la construcción de intervalos de confianza. Asimismo, se formalizan las cuestiones referentes a los contrastes de hipótesis.

En todos los casos, los problemas implicados se limitan al estudio de proporciones dentro de una población, y de medidas continuas que siguen una distribución normal.

Con algunas excepciones, los profesores de Enseñanza Secundaria no han abordado, ni en su formación ni en su práctica docente,

las cuestiones relacionadas con la estimación, ni con los test de contraste de hipótesis. Ello hace que no se conozcan los contenidos y las motivaciones que subyacen a este campo, por lo que difícilmente se podrán transmitir adecuadamente a los alumnos. En la práctica, se corre el riesgo de que la enseñanza de estos contenidos quede reducida a la marginalidad, e incluso que los profesores eludan en lo posible su impartición.

En nuestra opinión, este riesgo es particularmente grave por varios motivos:

1. Las pruebas de Selectividad incluyen ejercicios de Inferencia estadística.

2. En la mayor parte de las carreras universitarias de Ciencias humanas, sociales, jurídicas y naturales, existen asignaturas relacionadas con la Inferencia estadística, y en cualquier caso, las estimaciones y los contrastes de hipótesis están presentes en la enseñanza y en la práctica profesional.

3. La posibilidad de sugerir problemas relacionados con la vida real y, en concreto, con el entorno de los alumnos, hace que éstos tengan una importante motivación que ayuda a desarrollar estos temas, disminuyendo el riesgo de inhibición de los alumnos en la clase de Matemáticas. Adicionalmente, los temas transversales se pueden sugerir con facilidad como aplicaciones de la estimación y de los contrastes de hipótesis.

Como podrá observarse, la presentación de estos temas difiere de la utilizada en los anteriores, en la idea de que debe "formarse al profesor». Así, en el primero de ellos, dedicado a la estimación puntual y a la construcción de intervalos de confianza, se intenta que el profesor comprenda los fundamentos estadísticos del manejo de las muestras aleatorias. Obviamente, la unidad propuesta está muy formalizada, y no puede impartirse literalmente a los alumnos. Ahora bien, creemos que puede hacerse un recorrido de la misma menos árido, sustituyendo ciertas construcciones formales por intuiciones sobre lo que implica que una muestra sea representativa de la pobla-

ción en estudio. Al final, tal vez deban darse al alumno como una receta las expresiones de los intervalos de confianza, pero el profesor debe disponer de las ayudas a la interpretación que en el texto se presentan.

El segundo tema se dedica a la construcción de test de contrastes de hipótesis. Las técnicas empleadas son las clásicas (basadas en la teoría de Neyman y Pearson), y exigen la adquisición de un vocabulario (una "jerga») muy árido. Desgraciadamente, los elementos que participan en esta discusión son muchos y sin la ayuda de unas definiciones precisas, es fácil perderse. El texto se ha dividido en dos partes; en la primera se realiza un desarrollo formal, no exento de intuición, orientado al profesor. Sin su ayuda, ciertas cuestiones, como la fijación de niveles de significación o la falta de simetría entre hipótesis y alternativa, entre otras, pueden crear múltiples equívocos en su desarrollo en clase. La segunda parte del texto propone una unidad didáctica que, en nuestra opinión, puede desarrollarse sin excesivas dificultades en el tiempo de que se dispone para este tema. Como puede verse, en ella se prescinde de la mayor parte de los formalismos, pero es ineludible abordar (y formalizar parcialmente) la noción de riesgo (y su contraria, la confianza), si se quiere que el alumno tenga claro que las dificultades que se presentan no son gratuitas.

Como indicamos en la introducción de los materiales referentes a la ESO, publicados por el Instituto Superior de Formación del Profesorado en esta misma colección, no pretendemos realizar una programación ni un proyecto curricular, por lo que aquí también se obvian las consideraciones sobre objetivos y contenidos procedimentales y actitudinales, así como las referentes al proceso evaluatorio. Nos limitamos, por tanto, a exponer los contenidos conceptuales. Es cada profesor, en su entorno, quien definirá el resto de las cuestiones de acuerdo con su marco de trabajo.

El material didáctico se acompaña de algunas notas explicativas que sugieren referencias clásicas de los conceptos o lecturas propuestas que el profesor puede seguir si desea ampliar los contenidos aquí presentados. La mención a ciertos autores que se incluyen

en dichas notas y en los comentarios bibliográficos que se desarrollan al final de cada capítulo, se refieren a la bibliografía final que cierra el libro.

Qué duda cabe que la bibliografía no es completa, ni pretende serlo, sino que conscientemente se limita a materiales que se consideran de utilidad para alcanzar una preparación básica, obviando referencias redundantes o que se han considerado excesivamente arcaicas o especializadas.

En un anexo final se resumen diversas experiencias didácticas que los miembros del equipo han realizado con alumnos de sus centros. Estas experiencias se han venido desarrollando desde que los materiales se elaboraron hasta estos momentos, si bien su realización sistemática se abordó cuando los materiales fueron completados, es decir, hace más de seis años.

La distancia temporal con nuestros días dificulta su valoración, tanto por los cambios que se han producido en las programaciones de las disciplinas como por la diferencia entre los perfiles de los alumnos entre ambos períodos. Incluso las actuales disponibilidades de materiales multimedia y el hábito de su utilización convierten en áridos ciertos pasajes que en su día eran más innovadores.

Por ello, el resumen de la experimentación es conscientemente breve, limitándose a señalar las características y conclusiones que hoy se mantienen vigentes por depender de los contenidos o de su articulación temporal, y no de las precisiones expositivas.

1. VARIABLES ESTADÍSTICAS BIDIMENSIONALES. REGRESIÓN Y CORRELACIÓN LINEAL

1. INTRODUCCIÓN

Se ha realizado una encuesta a 40 alumnos, en la que se les preguntaba la nota que obtuvieron el curso pasado en las asignaturas de Física y Matemáticas. Los resultados aparecen a continuación:

Nota de Física	*Nota de Matemáticas*	*N.º de alumnos*
3	2	4
4	5	6
5	5	12
6	6	4
6	7	5
7	6	4
7	7	1
7	8	2
8	9	1
10	10	1

Los individuos que forman "la población" son los alumnos y "las características" que estamos estudiando son: "nota de Física", que denominaremos X, y "nota de Matemáticas", que denominaremos Y.

Al par formado por (X,Y) se le llama variable estadística bidimensional.

Lo primero que nos planteamos, ante un conjunto de datos referentes a una variable bidimensional, es buscar una manera de resumir la información. Una primera forma la constituyen las tablas de frecuencias cruzadas o "tablas de doble entrada" correspondientes al par (X,Y)

Para el ejemplo, la tabla de doble entrada sería la siguiente:

$X \mid Y$	2	5	6	7	8	9	10	$n_{i\bullet}$
3	4	0	0	0	0	0	0	4
4	0	6	0	0	0	0	0	6
5	0	12	0	0	0	0	0	12
6	0	0	4	5	0	0	0	9
7	0	0	4	1	2	0	0	7
8	0	0	0	0	0	1	0	1
10	0	0	0	0	0	0	1	1
$n_{\bullet j}$	4	18	8	6	2	1	1	40 = N

Observa que representamos por N el número total de individuos que forman la población objeto.

Prescindiendo, de momento, de la última fila y de la última columna, las casillas de la tabla recogen las frecuencias absolutas, que se definen a continuación:

La frecuencia absoluta del par de valores (x_i, y_j) es el número de individuos que presentan a la vez el valor x_i de X y el valor y_j de Y. La representaremos por n_{ij}.

Por ejemplo, n_{34} es la frecuencia absoluta del par (x_3, y_4), es decir, del par (5,7). Por lo tanto, $n_{34} = 0$, porque no hubo ningún alumno que sacara un 5 en Física y un 7 en Matemáticas. Análogamente, $n_{32} = 12$, porque hubo 12 alumnos que obtuvieron la tercera nota de Física $(x_3 = 5)$ y la segunda de Matemáticas $(y_2 = 5)$

La información contenida en la tabla de doble entrada, esto es, el conjunto formado por los valores de X e Y, y las frecuencias

absolutas, recibe el nombre de distribución bidimensional, o distribución conjunta, de X e Y.

La tabla de doble entrada puede contener, en lugar de las frecuencias absolutas, las frecuencias relativas, que son las frecuencias absolutas expresadas como proporción sobre el número total de individuos de la población, N.

Frecuencia relativa del par
$(x_i, y_j) = f_{ij} = n_{ij} / N$

Por ejemplo, f_{32} = 12/40 = 0,3, es decir, los alumnos que obtuvieron en las dos asignaturas un 5 fueron el 30% del total.

La suma de todas las frecuencias absolutas es N y, lo que es lo mismo, la suma de todas las frecuencias relativas es la unidad.

Comprueba la primera afirmación en la tabla anterior. Construye después la tabla de frecuencias relativas y comprueba la segunda afirmación. Observa que de las dos tablas podemos extraer la misma información, si conocemos el valor de N.

2. DISTRIBUCIONES MARGINALES

Consideremos ahora la última columna de la tabla de doble entrada aquí representada. Esa columna aparece en el margen derecho y contiene las frecuencias absolutas de la variable X, sin tener en cuenta la variable Y. Se les llama frecuencias absolutas marginales de X y se denotan por $n_{i\bullet}$.

Frecuencia absoluta marginal de $x_i = n_{i\bullet}$ = número de individuos que presentan el valor x_i de X independientemente del valor de Y que presenten. Es la suma de los valores de la fila correspondiente.

Por ejemplo, $n_{2\bullet} = 6$, es decir, el número de alumnos que obtuvieron la segunda nota de Física, o sea, un 4, fue 6.

Análogamente se definen las frecuencias relativas marginales:

Frecuencia relativa marginal de $x_i = f_{i\bullet} = n_{i\bullet} / N =$ proporción de individuos que presentan el valor x_i de X independientemente del valor de Y que presenten.

Si en vez de considerar la última columna, consideramos la última fila, que aparece en el margen inferior de la tabla, tendremos las frecuencias marginales (absolutas o relativas, dependiendo de lo que hayamos representado en la tabla) de Y. De forma que $n_{\bullet j}$ será la frecuencia absoluta marginal de y_j y $f_{\bullet j} = n_{\bullet j} / N$, será la frecuencia relativa marginal de y_j.

En definitiva, las columnas primera (que contiene los valores de la variable X) y última (que contiene las frecuencias marginales de X), o columnas marginales de la tabla, definen la distribución marginal de la variable X. De la misma forma, la información recogida en las filas marginales de la tabla constituye la distribución marginal de la variable Y.

3. DISTRIBUCIONES CONDICIONADAS

Hemos visto que el conjunto de valores y frecuencias de una variable estadística constituye "la distribución" de esa variable.

Vamos a definir ahora la distribución de una variable condicionada por un valor de otra (o condicionada a un valor de otra). Fijemos un valor para una de las dos variables de la distribución bidimensional, por ejemplo, $X = x_i$, y seleccionemos los valores que toma la variable Y en ese caso, es decir, cuando X toma el valor x_i. Tendremos entonces la variable Y condicionada por el valor x_i de X (se escribe $Y \mid X = x_i$). Los valores de esa variable y sus correspondientes frecuencias (que ya no son las frecuencias marginales de

Y) constituyen la distribución de Y condicionada por el valor x_i de X (o condicionada a que X toma el valor x_i).

Las frecuencias absolutas de esta distribución serán los n_{ij} cuando i permanece fijo. Por ejemplo, en la distribución de $Y \mid X = x_1$, las frecuencias absolutas serán $n_{11}, n_{12}, n_{13}, \ldots$ y se cumplirá que la suma de todas ellas será igual al número de individuos que presentan el valor x_1 de X, es decir, $n_{1\bullet}$. En general:

Frecuencias absolutas de la distribución $Y \mid X = x_i$:

$$n_{i1}, n_{i2}, n_{i3}, \ldots; \qquad \sum_{j=1}^{p} n_{ij} = n_{i\bullet};$$

siendo p el número de valores distintos de Y.

Las frecuencias relativas de la distribución de $Y \mid X = x_i$ serán lógicamente los cocientes entre las absolutas y la suma de todas ellas, es decir:

$$f\left(Y = y_j \mid X = x_i\right) = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}N}{f_{i\bullet}N} = \frac{f_{ij}}{f_{i\bullet}}$$

De ahí que la frecuencia relativa condicionada puede también calcularse como el cociente entre las frecuencias relativas conjunta y marginal.

Observa que bajo las fórmulas, lo único que subyace es la idea de que si nos quedamos exclusivamente con los individuos para los que la variable X toma el valor x_i, y prescindimos de los demás, en primer lugar, la proporción de los que presentan el valor y_j de Y se calculará como cualquier proporción, y, en segundo lugar, esa proporción podrá ser obtenida como cociente de dos proporciones.

De la relación anterior se deduce inmediatamente que la frecuencia relativa conjunta del par (x_i, y_j) es el producto de la frecuencia relativa marginal de x_i y la frecuencia relativa de y_j condicionada al valor x_i:

$$f_{ij} = f_{i\bullet} f\left(Y = y_j \mid X = x_i\right)$$

Observa nuevamente que las fórmulas están recogiendo ideas obvias, pues si la relación anterior la expresamos en función de las frecuencias absolutas, obtenemos una identidad:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n}\,\frac{n_{ij}}{n_{i\bullet}}$$

Hemos trabajado con la distribución de $Y \mid X = x_i$. Igualmente podemos hacerlo con la distribución de $X \mid Y = y_j$, sin más que considerar fijo un valor de Y, es decir, seleccionando a los individuos que presentan un determinado valor de Y y prescindiendo de los demás.

En este caso, las fórmulas anteriores serán:

Frecuencias absoluta de la distribución $X \mid Y = y_j$:

$$n_{1j}, n_{2j}, n_{3j}, \ldots; \qquad \sum_{i=1}^{k} n_{ij} = n_{\bullet j},$$

siendo k el número de valores distintos de X.

$$f\left(X = x_i \mid Y = y_j\right) = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij} N}{f_{\bullet j} N} = \frac{f_{ij}}{f_{\bullet j}}$$

$$f_{ij} = f_{\bullet j} f\left(X = x_i \mid Y = y_j\right)$$

En el ejercicio que estamos utilizando, obtén la distribución de la variable "nota de Matemáticas" condicionada a que la "nota de Física" fue un 7. Comprueba en esa distribución todas las relaciones anteriores. Después, responde a la pregunta: ¿Cuántas distribuciones unidimensionales es posible obtener de esta distribución bidimensional (nota de Física, nota de Matemáticas)? ¿Y en cualquier distribución bidimensional?

4. REPRESENTACIONES GRÁFICAS

La representación gráfica más útil para una variable bidimensional es el diagrama de dispersión o nube de puntos, que se obtiene representando cada par de observaciones (x_i, y_j) mediante un punto del plano. Cuando la frecuencia absoluta es mayor que uno (el par aparece en la población más de una vez),se puede representar mediante un punto cuyo grosor esté relacionado con el valor de dicha frecuencia. Pero resulta mucho más claro ofrecer, junto al gráfico, las correspondientes frecuencias.

El diagrama de dispersión permite visualizar la posible relación entre las dos variables.

Si representamos en el dibujo un nuevo par de ejes de coordenadas con centro en el punto $(\bar{x}, \bar{y})$, podemos obtener gráficos del tipo siguiente:

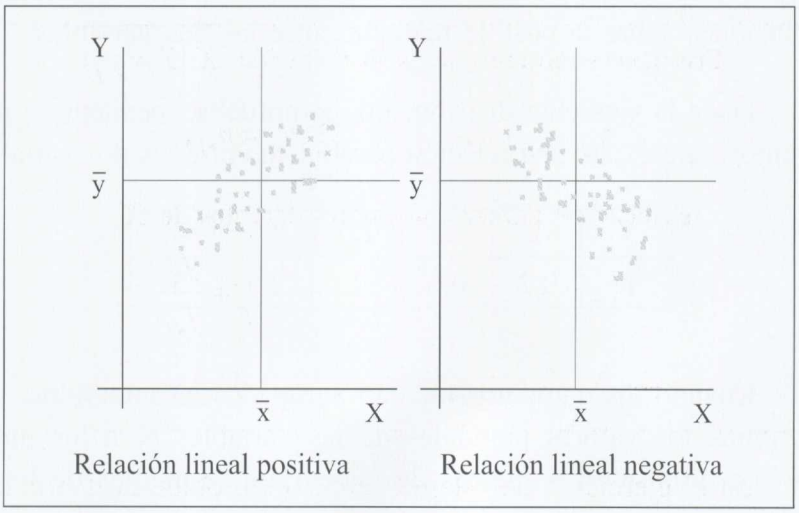

Figura 1.1

En el primer diagrama, se observa que los puntos se distribuyen, más o menos, en torno a una línea recta. Además, se trata de una recta de pendiente positiva (en términos generales, los incrementos en el valor de una de las variables se corresponden con incrementos en el valor de la otra, o sea, las dos variables se mueven en el mismo sentido). Todo ello sugiere lo que denominaremos una relación lineal "positiva" entre las dos variables.

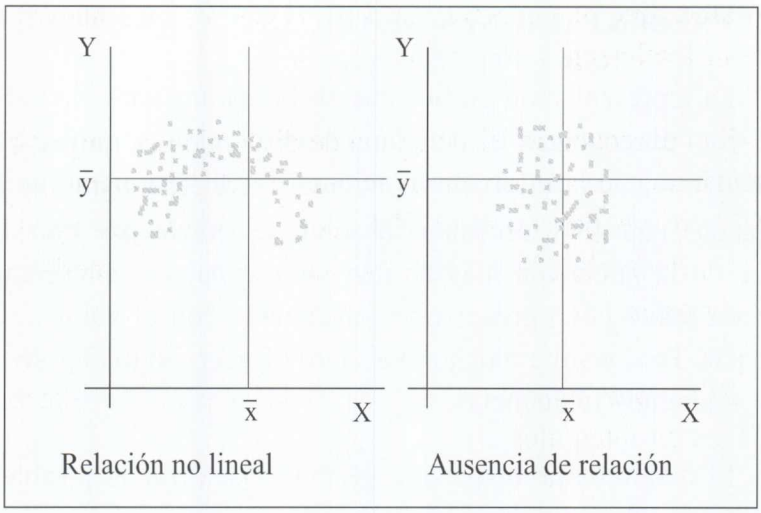

Figura 1.2

De forma análoga podríamos razonar en el resto de los casos.

Representa los dos diagramas de dispersión de nuestro ejemplo (Y en función de X y X en función de Y) y saca tus propias conclusiones sobre la posible relación entre las dos variables.

Dada la siguiente distribución, comprueba, mediante la nube de puntos, que existe una relación parabólica entre las dos variables:

X	1	2	2,5	3,5	4
Y	0,2	0,6	1	2,5	5

Cuando los datos nos los dan agrupados en intervalos, y no conocemos los valores puntuales de las variables para los individuos, recuerda que las marcas de clase (puntos medios de los intervalos) actúan como representantes de esos valores, de forma que sería posible construir el diagrama de dispersión representando dichas marcas.

El problema es que hay que poner muy en duda la representatividad de una nube de puntos así obtenida, porque, al no ser realmente la de la distribución bidimensional (X,Y), sino la de otra en la que se asume que todos los individuos de una clase presentan el mismo valor de la variable correspondiente, puede suceder que las

conclusiones que obtengamos, siendo válidas para las nuevas variables, sean totalmente erróneas para (X,Y).

En consecuencia, el diagrama de dispersión es muy útil cuando disponemos de la información puntual o concreta para los individuos, y, aunque para otro fin del estudio, decidamos agrupar los valores en intervalos, la nube de puntos debemos obtenerla con los datos originales.

En lo que sigue, nunca consideraremos que los datos están expresados en intervalos.

5. INDEPENDENCIA

Se dice que la variable Y es independiente de la variable X si, para todo valor de Y, y_j, la frecuencia de y_j condicionada a cualquier valor de X, es igual a la frecuencia marginal de y_j:

$$\forall y_j \quad f(Y = y_j \mid X = x_1) = f(Y = y_j \mid X = x_2) = \ldots = f_{\bullet j}$$

Por ejemplo, diremos que la nota de Inglés es independiente de la nota de Matemáticas, si la proporción de individuos que han sacado un 7 en Inglés, entre los que han sacado un 5 en Matemáticas, coincide con la proporción de los que han obtenido un 7 en Inglés, entre los que han sacado un 2 en Matemáticas, y con la proporción de los que han obtenido un 7 en Inglés, si sólo cogemos a los que han sacado un 0 en Matemáticas,..., y también coincide con la proporción de individuos que han obtenido un 7 en Inglés dentro de la población total de alumnos. Pero es más, lo mismo podríamos afirmar cambiando el 7 de Inglés por cualquier otra nota (en la definición se dice "para todo valor de Y"). Es evidente que, si eso es así, la nota de Inglés no tiene nada que ver con la nota de Matemáticas.

Si ahora consideramos la relación que existe siempre entre frecuencia conjunta, frecuencia marginal de x_i y frecuencia de y_j condicionada a x_i, en este caso se tiene:

Si Y es independiente de X, la frecuencia relativa conjunta es el producto de las marginales.

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

Es fácil demostrar que la independencia siempre es recíproca; es decir, que si Y es independiente de X, X también lo es de Y, y, por tanto:

$$\forall x_i \qquad f(X = x_i \mid Y = y_1) = f(X = x_i \mid Y = y_2) = \ldots = f_{i\bullet}$$

En el caso de independencia, la relación entre las frecuencias absolutas (se deduce de la que existe entre las relativas) es:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}$$

Sobre la tabla de doble entrada es muy fácil analizar la independencia, porque, en ese caso, todas las columnas (incluyendo la del margen derecho) son proporcionales entre sí. Lógicamente ocurre lo mismo con las filas, incluyendo la última.

6. ASOCIACIÓN LINEAL. LA COVARIANZA Y EL COEFICIENTE DE CORRELACIÓN

La covarianza es una medida del grado de asociación lineal que existe entre dos variables y se define como:

$$S_{XY} = \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{p} (x_1 - \bar{x})(y_j - \bar{y}) n_{ij}$$

Por comodidad, en lugar de trabajar con f_{ij} o con n_{ij}, se puede actuar de la siguiente forma: Consideramos todos los valores de $[(x_i - \bar{x})(y_j - \bar{y})]$, no sólo los distintos, repitiendo aquellos que aparezcan más de una vez tantas veces como aparezcan (tantas veces

como indique n_{ij}). Entonces, ya no habrá que multiplicar por n_{ij}. Así que la fórmula de la covarianza será:

$$S_{XY} = \frac{1}{N}\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})$$

A partir de la expresión anterior, se puede obtener otra que, en algunos casos, resulta muy útil:

$$\begin{aligned} S_{XY} &= \frac{1}{N}\sum_{t=1}^{N}(x_t y_t - x_t\bar{y} - \bar{x}y_t + \overline{xy}) \\ &= \frac{1}{N}\left(\sum_{t=1}^{N} x_t y_t - \bar{y}\sum_{t=1}^{N} x_t - \bar{x}\sum_{t=1}^{N} y_t + N\,\overline{xy}\right) \\ &= \frac{\sum_{t=1}^{N} x_t y_t}{N} - \bar{y}\frac{\sum_{t=1}^{N} x_t}{N} - \bar{x}\frac{\sum_{t=1}^{N} y_t}{N} + \overline{xy} \\ &= \frac{\sum_{t=1}^{N} x_t y_t}{N} - \overline{yx} - \overline{xy} + \overline{xy} \\ &= \frac{1}{N}\sum_{t=1}^{N} x_t y_t - \overline{xy} \end{aligned}$$

Por tanto, la covarianza también es la diferencia entre la media del producto de las dos variables y el producto de las medias.

La palabra "covarianza" tiene que ver con la expresión "variación conjunta", porque su objetivo es precisamente medir la variación conjunta de las dos variables. Si nos fijamos en las figuras del apartado "representaciones gráficas", podemos observar lo siguiente:

- Cuando hay una *relación lineal positiva*, los puntos del diagrama de dispersión se encuentran situados, en relación con los nuevos ejes de coordenadas, en los cuadrantes primero y tercero, es decir, $(x_i - \bar{x})$ e $(y_j - \bar{y})$ tienen el mismo signo, por tanto, la covarianza es positiva y relativamente alta.

- Cuando hay una *relación lineal negativa*, los puntos están en los cuadrantes segundo y cuarto, es decir, $(x_i - \bar{x})$ e $(y_j - \bar{y})$ tienen distinto signo, por tanto, la covarianza es negativa y, en valor absoluto, relativamente alta.
- Cuando *apenas existe relación lineal*, la covarianza, sea positiva o negativa, es, en valor absoluto, relativamente pequeña (relativamente próxima a cero)

Además, se demuestra que si dos variables son independientes, la covarianza entre ambas es nula. Sin embargo, lo contrario no es cierto, es decir, el que la covarianza entre dos variables sea cero sólo significa que el grado de dependencia lineal es nulo, pudiendo existir otro tipo de relación entre ambas.

El inconveniente de la covarianza es que su valor depende de las unidades de medida de las dos variables, de manera que si, por ejemplo, calculamos la covarianza entre el peso expresado en kilogramos y la altura expresada en centímetros, de ciertos individuos, el valor obtenido será cien veces mayor que el que obtendríamos expresando la altura en metros (y el peso en kilogramos). De esto se deriva que la covarianza no está acotada. Por lo tanto, no es una medida muy adecuada para comparar distintos grados de asociación lineal (salvando el caso en que las parejas de variables estén expresadas en las mismas unidades y sólo nos interese ordenar el grado de relación lineal)

El problema anterior lo resuelve la medida que veremos a continuación: *El Coeficiente de Correlación Lineal*[1].

Se define el coeficiente de correlación lineal entre dos variables como el cociente entre la covarianza y el producto de las desviaciones típicas:

1. Galton (1822-1911) introdujo el concepto de correlación en sus investigaciones sobre identificación de criminales (véase, por ejemplo, Stigler (STIGLER, S.M. *The History of Statistics. The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University Press. Cambridge, Mass., 1986. Pág. 265), en las que analizaba las relaciones entre ciertas características antropométricas, como la altura, la longitud del antebrazo, etc. Los estudios de correlación fueron también llevados a cabo por Karl Pearson (1857-1930), quien trabajaba con Galton (STIGLER, S.M. Opus cit. 1986. Pág. 326). De hecho, éste es el motivo por el que a este coeficiente se le conoce como coeficiente de correlación lineal de Pearson.

$$r = \frac{S_{XY}}{S_X S_Y},$$

donde

$$S_{XY} = \frac{1}{N}\sum_{t=1}^{N} x_t y_t - \overline{xy}$$

$$S_X^2 = \frac{1}{N}\sum_{t=1}^{N}(x_t - \overline{x})^2 = \frac{1}{N}\sum_{t=1}^{N} x_t^2 - \overline{x}^2$$

$$S_Y^2 = \frac{1}{N}\sum_{t=1}^{N}(y_t - \overline{y})^2 = \frac{1}{N}\sum_{t=1}^{N} y_t^2 - \overline{y}^2$$

Esta medida, que, al igual que la covarianza, sirve para evaluar el grado de asociación lineal existente entre dos variables, posee las siguientes *propiedades*:

1. Su valor está comprendido entre -1 y 1:

$$-1 \le r \le 1$$

2. Los dos valores extremos corresponden al caso de relación lineal exacta entre las variables. Es decir:

$$Y = aX + b \Leftrightarrow \begin{cases} r = 1, & \textbf{si } a > 0 \\ r = -1, & \textbf{si } a < 0 \end{cases}$$

3. Si dos variables son independientes, el coeficiente de correlación lineal entre ellas es cero. Lo contrario no es cierto: si el coeficiente de correlación es nulo, entre las dos variables no existe relación lineal, pero puede existir otro tipo de relación.

   Por ejemplo, dada la variable $X = \{-2,-1,0,1,2\}$, puedes construir la variable $Y = X^2$. Evidentemente, cada una de las variables es una función no lineal perfecta de la otra. Sin embargo, el coeficiente de correlación lineal

entre ambas es cero, porque no existe relación lineal entre ellas. Compruébalo.

4. Es una medida invariante a cambios de escala y/u origen, así que no depende de las unidades de medida en que vengan expresadas las variables. Es decir, que si multiplicamos X e Y por dos constantes, y/o sumamos dos constantes a X e Y, el coeficiente de correlación no varía.

El significado y la interpretación del coeficiente de correlación lineal se derivan de las propiedades anteriores.

En el ejemplo de las notas de Física y de Matemáticas, se obtienen los resultados $S_{XY} = 2{,}6$ **y** $r_{XY} = 0{,}92$. Mientras que el valor de la covarianza sólo nos dice que la correlación lineal entre las dos variables es positiva, el valor de r indica:

1. Que existe una fuerte correlación lineal entre las dos notas, pues el valor absoluto de r está próximo a la unidad.

2. Que la correlación es positiva, pues r es positivo. Eso significa que las dos notas se mueven en el mismo sentido (en general, si un alumno tiene mayor nota en Física que otro, también tendrá mayor nota en Matemáticas)

La fuerte relación lineal entre dos variables no implica relación de causalidad (causa-efecto) entre ellas. Si el número de cigüeñas que anidan en el campanario de la iglesia de una serie de pueblos y el número de nacimientos producidos en los mismos pueblos están fuertemente correlacionados ¿te atreverías a afirmar que la afluencia de cigüeñas aumenta el número de nacimientos? ¿o asegurarías, tal vez, que a los niños los trae la cigüeña?[2].

2. Los estadísticos advierten frecuentemente sobre los peligros del abuso de la estadística, con ejemplos en los que una fuerte correlación entre dos variables no debe interpretarse como una relación causal, sino como un efecto de la casualidad o como la existencia de terceros factores encubiertos que influyen en las dos variables, provocando la alta correlación. Ver, por ejemplo, PEÑA, Daniel y ROMO, Juan. *Introducción a la estadística para las ciencias sociales.* McGraw-Hill. Madrid, 1997.

Analiza, también, la siguiente afirmación ficticia: "Dado que existe una fuerte correlación entre el consumo de zapatos y el consumo de libros, para fomentar la lectura en la población, el gobierno debería subvencionar la compra de zapatos".

Desde luego, existen muchos casos en que, posiblemente, la fuerte correlación está acompañada de una relación de causalidad. Por ejemplo, si el gasto familiar en bienes de lujo está fuertemente correlacionado con los ingresos familiares, ¿crees que también existe una relación de causalidad en algún sentido? ¿o la correlación es puramente casual?

7. RECTAS DE REGRESIÓN

7.1. Recta de regresión de Y sobre X

Dadas dos variables, podemos asumir, en principio, que cada una es, más o menos aproximadamente, una función lineal de la otra. Si realmente existe una relación lineal perfecta entre ellas, el grado de aproximación será total y si lo que ocurre es que no hay relación lineal alguna, el grado de aproximación será nulo.

Asumamos, pues, que, aproximadamente, Y es una función lineal de X. Expresaremos Y como una función lineal exacta de X, $Y = aX + b$, y, posteriormente, evaluaremos el grado de aproximación, es decir, lo poco o mucho que nos hemos confundido al suponer una relación lineal perfecta (más adelante, utilizaremos la expresión "bondad de ajuste").

Si la relación supuesta fuera cierta, y conociéramos los valores de a y b, podríamos, entre otras cosas, predecir el valor de Y, a partir del valor conocido de X.

La ecuación $Y = aX + b$, una vez que hayamos obtenido, según un determinado criterio, los valores de a y b (la pendiente y la

ordenada en el origen), recibe el nombre de recta de regresión de Y sobre X.

Para determinar los valores de a y b, se utiliza el criterio de los mínimos cuadrados[3], cuyo razonamiento veremos a continuación.

En la figura 1.3 hemos representado el diagrama de dispersión o la nube de puntos de Y frente a X (Y en ordenadas, X en abscisas). Nos gustaría encontrar una recta que explicara lo mejor posible el comportamiento de Y en función de X, es decir, que se

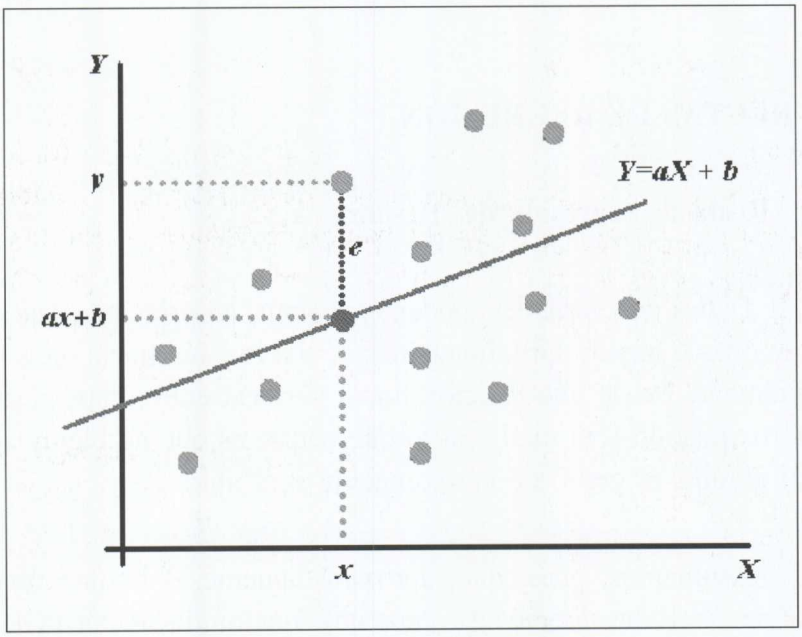

Figura 1.3

3. El criterio de los mínimos cuadrados fue introducido en 1795 por Carl Friedrich Gauss (véase, por ejemplo, Stigler (STIGLER, S.M. Opus cit. 1986. Pág. 140), quien lo aplicó al estudio de los errores en las mediciones astronómicas y, de forma independiente, por Adrien Marie Legendre, en 1805 (SMITH, D.E. *Legendre on least squares*, in *A source book of mathematics*. McGraw-Hill. New York, 1929. Reimpresión de Ed. Dover. New York, 1959. Págs. 576 a 579, describe detenidamente el trabajo de este matemático sobre el problema de mínimos cuadrados). Posteriormente, Francis Galton (1822-1911) realiza grandes aportaciones a la estadística, con sus estudios de regresión y correlación, con los que pretendía contrastar las teorías evolucionistas de Charles Darwin, primo suyo, publicadas en *El origen de las especies,* en 1859. Galton observó que el tamaño de las semillas de plantas de guisantes hijas "revertía" al tamaño medio, así como que la estatura de los hijos "regresa" hacia el valor medio (GALTON, F: *Natural Inheritance*. McMillan. Londres, 1889). De ahí surgió el nombre de la técnica de Regresión.

ajustara lo mejor posible a la nube de puntos. Esa será la recta de regresión de Y sobre X, por el momento desconocida. Supongamos que la conocemos y representémosla en el mismo gráfico (en realidad, hemos representado una recta cualquiera). Observemos un valor cualquiera de la variable X, por ejemplo x. Para ese valor de X, el valor observado o real de Y es y. Sin embargo, según la recta de regresión, la ordenada correspondiente a ese valor de abscisas es $ax+b$, o sea, ese es el valor estimado o ajustado de Y, para el valor x de X. La diferencia entre el valor observado y el valor estimado se llama error o residuo correspondiente a x, y lo denotaremos por e.

Para cada valor de X, cometemos un error, al sustituir el verdadero valor de Y por el estimado. Existen tantos errores como individuos, es decir, N. Si todos los errores fuesen nulos (o, lo que es lo mismo, si los valores estimados y observados de Y coincidiesen) la recta pasaría por todos los puntos de la nube y, por tanto, el ajuste sería perfecto. Por el contrario, cuanto mayores, en valor absoluto, sean los errores (cuanto más alejados estén los valores estimados y observados de Y), peor será el ajuste realizado.

El criterio de los mínimos cuadrados consiste en elegir, entre todas las rectas posibles, aquélla para la cual la suma de los cuadrados de los errores sea mínima. Dicho de otra forma, no existe otra recta que proporcione una suma de cuadrados de los errores más pequeña.

Los valores de a y b que se obtienen aplicando el criterio anterior son los siguientes:

$$a=\frac{S_{XY}}{S_X^2}=r\frac{S_Y}{S_X};\qquad b=\bar{y}-a\bar{x}$$

Por tanto, la ecuación de la recta de regresión de Y sobre X es:

$$Y=\frac{S_{XY}}{S_X^2}X+\left(\bar{y}-a\bar{x}\right)=\frac{S_{XY}}{S_X^2}X+\bar{y}-\frac{S_{XY}}{S_X^2}\bar{x}$$

o, lo que es lo mismo:

$$Y - \bar{y} = \frac{S_{XY}}{S_X^2}(X - \bar{x})$$

7.2. Interpretación de la pendiente en la recta de Y sobre X

Teniendo en cuenta el significado de la pendiente de una recta, la interpretación de a en la recta de Y sobre X es la siguiente:

Cuando X aumenta en una unidad, Y lo hace en a unidades. Si el valor de a es negativo, por ejemplo -3, diremos que cuando X aumenta en una unidad, Y disminuye en tres unidades (en lugar de decir que Y aumenta en menos tres unidades, que vendría a significar lo mismo).

En definitiva, la pendiente mide la influencia sobre Y de un cambio unitario en X.

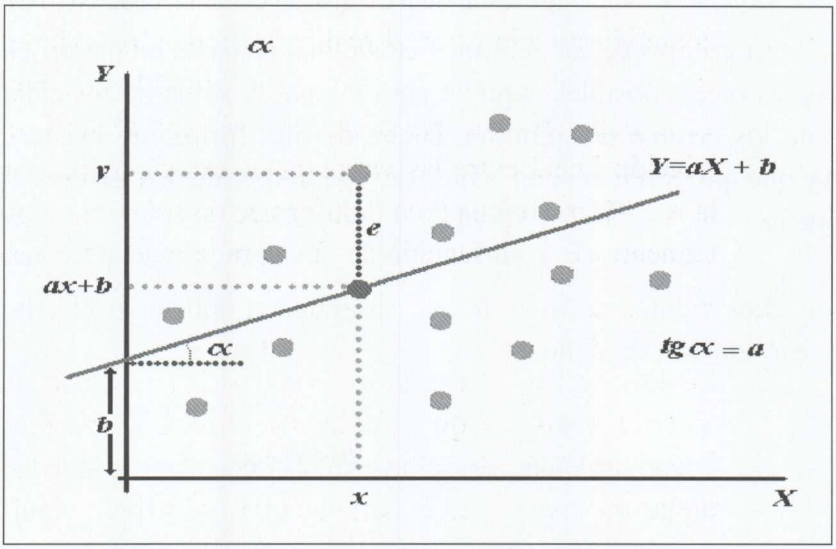

Figura 1.4

7.3. Bondad del ajuste en la recta de regresión de Y sobre X

Para analizar la bondad del ajuste, en la regresión con dos variables, que es nuestro caso, podemos basarnos en el cuadrado del coeficiente de correlación lineal.

Teniendo en cuenta que $0 \leq r^2 \leq 1$, la interpretación de esta medida es la siguiente:

1. Si $r^2 = 0$, significa que no existe tal relación lineal entre las variables (puede existir otro tipo de relación o no haber ninguna entre las dos variables). La recta es la que, según el criterio de los mínimos cuadrados, mejor se ajusta a los puntos de la nube; aun así, la bondad del ajuste es nula, porque Y no es una función lineal de X. La recta no proporciona ninguna información sobre el comportamiento de Y en función de X y no tendría sentido utilizarla para analizar la influencia de X sobre Y, ni para predecir el valor de Y, dado X.

2. Si $r^2 = 1$, significa que el ajuste es perfecto. No hemos cometido ningún error al realizarlo (todos los valores de los errores son nulos). La recta pasa por todos los puntos de la nube, obviamente porque éstos están alineados. La relación lineal entre las variables es exacta, y, por tanto, la recta proporciona toda la información sobre el comportamiento de Y en función de X, para la muestra considerada.

3. Si $0 < r^2 < 1$, en función de a cual de los dos situaciones anteriores nos acerquemos, hablaremos de ajuste malo o bueno. Además, el valor concreto de r^2 se puede interpretar en los siguientes términos: si $r^2 = 0{,}85$, significa que la recta obtenida explica en un 85% el comportamiento de Y en función de X, y el 15% restante viene explicado por los errores cometidos.

7.4. Predicción en la regresión de Y sobre X

Predecir el valor de Y es determinar el valor de dicha variable, suponiendo conocido el valor de X. La predicción se realiza simplemente sustituyendo el valor de X en la ecuación de la recta de regresión de Y sobre X y hallando el valor resultante de Y.

Para que la predicción sea fiable, es necesario que se cumplan, al menos, las siguientes condiciones:

- Que el valor de X sea correcto, o, si es un valor futuro, que exista una relativa confianza en su fiabilidad.
- Que el grado de bondad del ajuste realizado sea alto (r^2 próximo a uno).
- Que la predicción no se efectúe en otras condiciones, pues la recta ha sido obtenida en base a unos datos concretos y no hay garantía de que sea válida para otros.

7.5. Recta de regresión de X sobre Y

La recta de regresión de X sobre Y se obtiene de forma análoga a la de Y sobre X.

Representa el diagrama de dispersión de X frente a Y (puedes representar X en ordenadas e Y en abscisas) e intenta realizar el razonamiento que hicimos en el caso anterior. Llama ahora recta de regresión de X sobre Y a la ecuación $X = a'Y + b'$.

Los resultados que se obtienen en este caso son los siguientes:

$$a' = \frac{S_{XY}}{S_Y^2} = r\frac{S_X}{S_Y}; \qquad b' = \bar{x} - a'\bar{y}$$

$$X = \frac{S_{XY}}{S_Y^2}Y + (\bar{x} - a'\bar{y}) = \frac{S_{XY}}{S_Y^2}Y + \bar{x} - \frac{S_{XY}}{S_Y^2}\bar{y}$$

$$(X-\bar{x})=\frac{S_{XY}}{S_Y^2}(Y-\bar{y})$$

Observa que la recta de regresión de X sobre Y no se obtiene despejando X en la recta de regresión de Y sobre X. Si hacemos esto último, seguiremos teniendo la ecuación de la recta de Y sobre X, sin que importe cuál de las dos variables hemos despejado. En general, las dos rectas son diferentes.

La medida de bondad de ajuste, en este caso, sigue siendo el cuadrado del coeficiente de correlación lineal. Así que no tiene sentido plantearse qué recta de las dos proporciona un ajuste mejor. Cuando se trabaja con datos reales, es importante tener claro el sentido de la causalidad (qué variable produce más claramente efectos sobre la otra), para seleccionar una de las dos. Es decir, en realidad, no se predicen simultáneamente Y en base a X (a partir de la recta de Y sobre X) y X en base a Y (a partir de la recta de X sobre Y).

7.6. Comparación de las dos rectas de regresión

1. El producto de las dos pendientes es el cuadrado del coeficiente de correlación lineal:

$$aa' = r^2$$

2. Las dos rectas son distintas, salvo en el caso de ajuste perfecto en que coinciden: si $r^2 = 1$, la recta de X sobre Y se obtiene despejando X en la recta de Y sobre X, es decir, se trata de la misma ecuación.

3. En lo que sigue, para comparar mejor las dos rectas, vamos a suponer que las representamos juntas, en el mismo sistema de ejes cartesianos, con X en abscisas e Y en ordenadas. Entonces:

    a) Las dos rectas se cortan en el punto $(\bar{x}, \bar{y})$.
    b) Las dos tienen pendiente, bien positiva (cuando r es positivo), bien negativa (cuando r es negativo).

c) La recta de X sobre Y es la que tiene mayor pendiente en valor absoluto.

d) Si el coeficiente de correlación es nulo, las dos rectas son paralelas a los ejes de coordenadas y perpendiculares entre sí.

Los dibujos siguientes ilustran estas características.

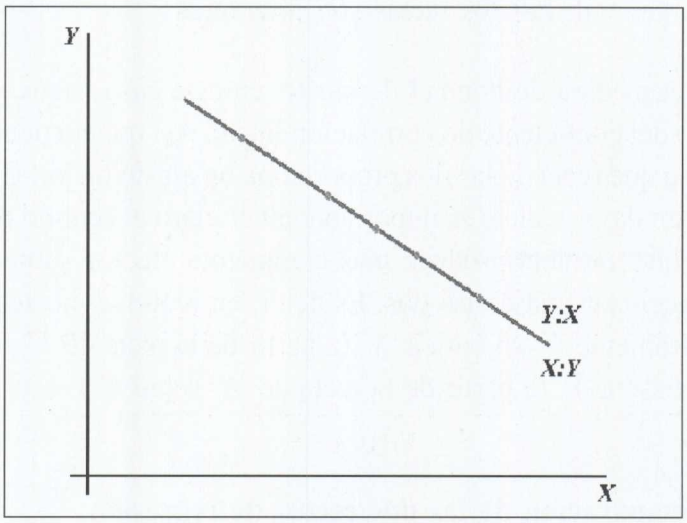

Figura 1.5.a

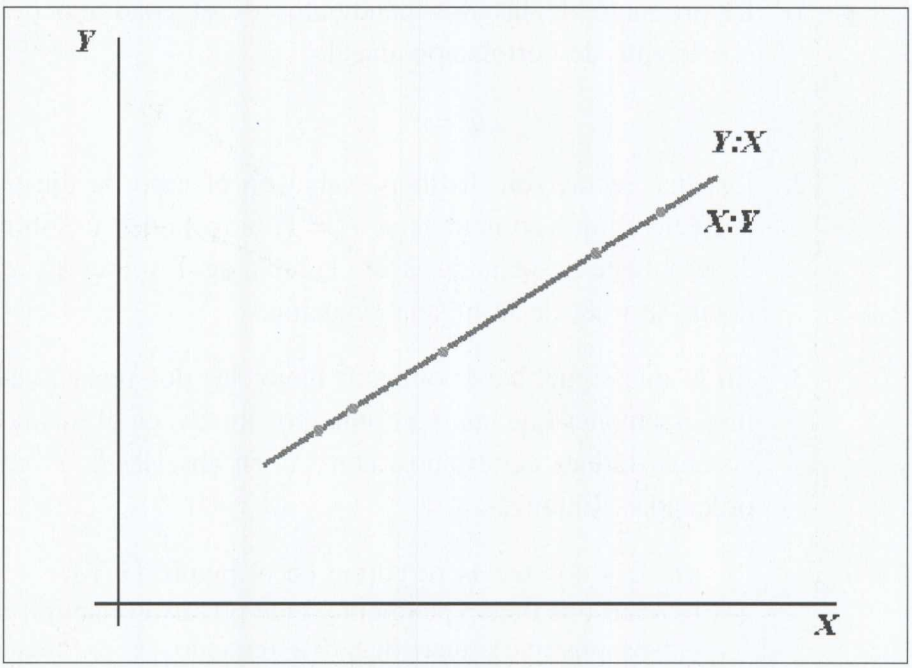

Figura 1.5.b

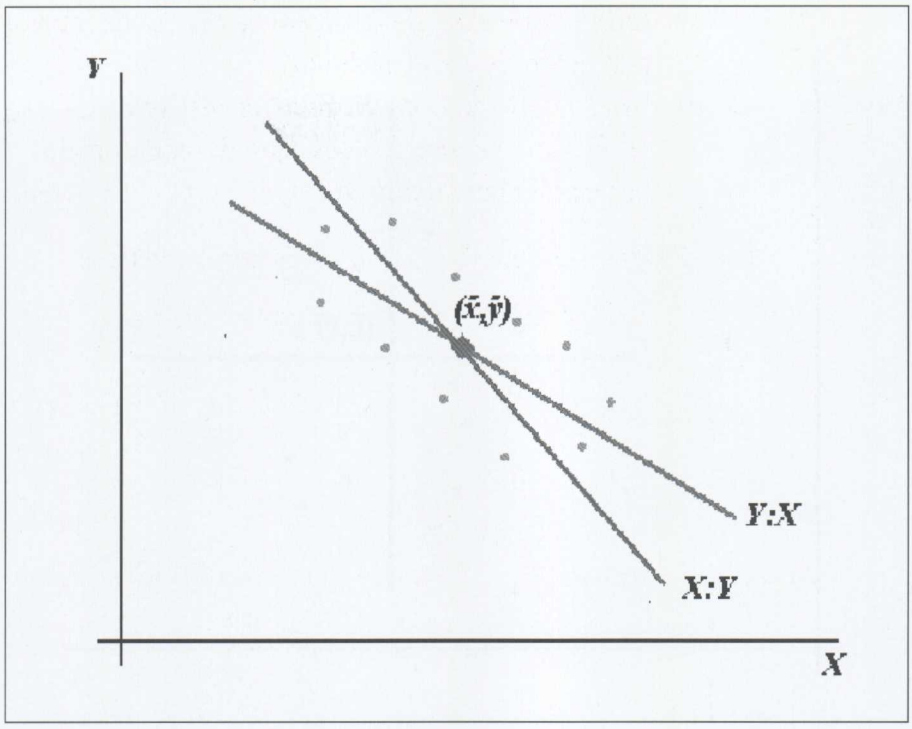

Figura 1.5.c

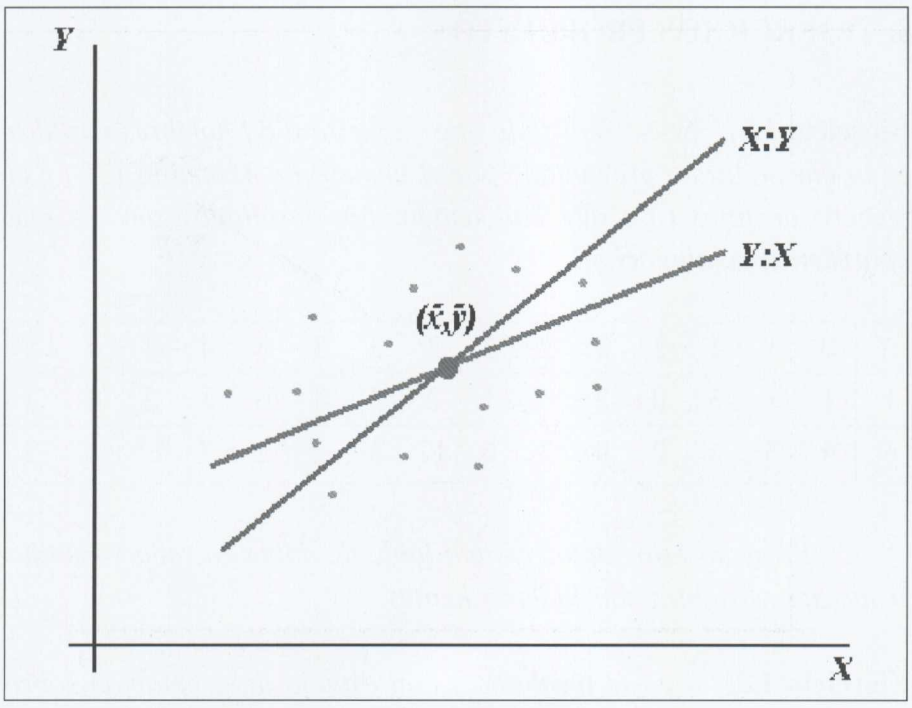

Figura 1.5.d

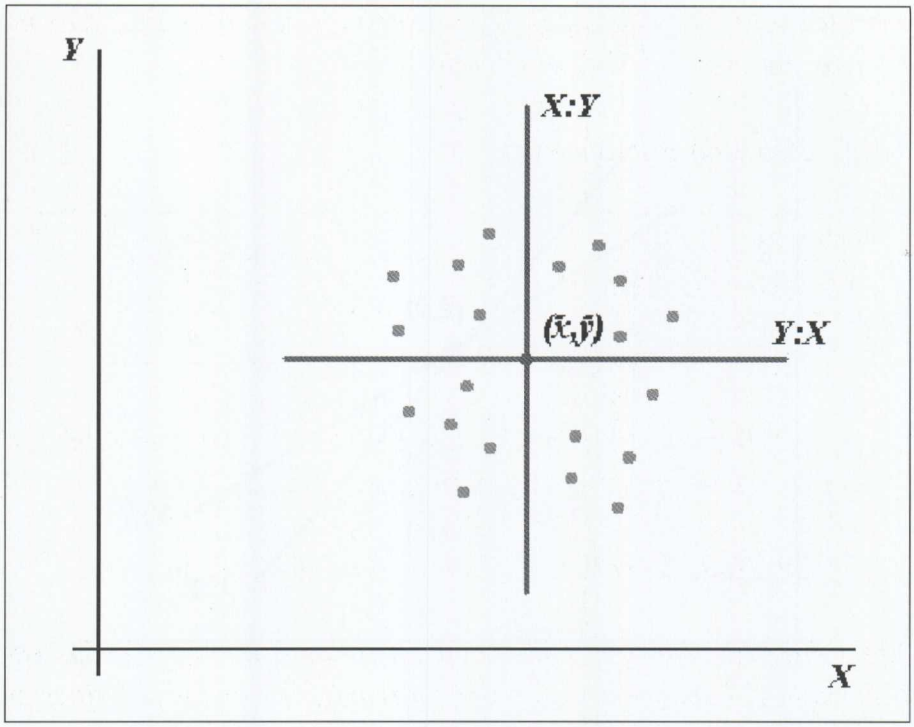

Figura 1.5.e

8. EJERCICIOS PROPUESTOS

Ejercicio 1.1. *Se ha realizado una encuesta a 80 hombres casados en la que se les ha preguntado por el número de hermanos (X) y el número de hijos (Y) que han tenido. Los resultados son los que aparecen a continuación.*

X	0	1	1	2	2	2	3	3	3	3	4	4	4	4	5
Y	1	1	3	0	2	3	1	2	3	4	0	1	2	4	2
n_{ij}	4	3	4	2	9	3	6	12	5	2	2	7	15	1	5

Presenta estos datos en una tabla de doble entrada y estudia si las dos variables son independientes.

Ejercicio 1.2. *Con el fin de hacer un estudio de aceptación sobre dos modelos de automóviles de reciente fabricación, se han conside-*

rado las ventas efectuadas por un concesionario durante los días no festivos del mes de septiembre último.

Han sido las siguientes:

Ventas Modelo A	0	1	2	2	3	3	4	4
Ventas Modelo B	2	3	1	2	1	2	0	1
Nº de días	1	1	3	5	8	4	1	2

1. *Halla las distribuciones marginales correspondientes, así como sus medias y varianzas.*

2. *Halla la covarianza de la distribución dada.*

Ejercicio 1.3. *La tabla que se da a continuación describe las calificaciones obtenidas en la prueba teórica (X) y en las prácticas de laboratorio (Y) por los alumnos de la asignatura de Química de una Escuela Universitaria.*

X	3	4	4	4	4	4	5	5	5	5	5	5	6	6	6	6	7
Y	4	4	4	4	5	5	5	5	6	6	6	7	7	7	7	7	6

X	7	7	7	7	7	7	7	7	8	8	8	8	9	9	9	9	10
Y	7	7	7	7	7	7	8	8	7	7	7	8	7	8	8	9	9

1. *Obtén la tabla de doble entrada con la distribución conjunta de Frecuencias.*

2. *Halla las distribuciones marginales correspondientes; sus medias y sus varianzas.*

3. *Calcula la covarianza de la distribución bidimensional.*

4. *Calcula la media de X condicionada a que $Y = 7$.*

5. *Calcula la varianza de Y condicionada a que $X = 5$.*

Ejercicio 1.4. *Dada la distribución estadística:*

X	3	3	4	4	5
Y	8	9	10	11	11

1. *Dibuja el diagrama de dispersión de Y frente a X. ¿Qué tipo de relación sugiere?*
2. *Obtén las dos rectas de regresión.*

Ejercicio 1.5. *Consultado el fichero de un departamento de pediatría, se obtuvieron los siguientes datos respecto al peso y a la edad de los niños atendidos:*

Peso en Kg\ Edad	*0*	*1*	*2*	*3*	*4*
0-5	2				
5-10	4	2			
10-15		8	9	7	
15-20		1	2	8	14
20-25					1

1. *Halla el coeficiente de correlación.*
2. *Obtén la recta de regresión de Y (peso) sobre X (edad).*
3. *Con la recta obtenida, decide cuál es el peso que debe esperarse para un niño de 5 años.*

Ejercicio 1.6. *La tabla que se da a continuación representa los ingresos totales y los gastos fijos por mes, en euros, de un grupo de 40 familias. Calcula el coeficiente de correlación y la recta de regresión de los gastos fijos (Y) sobre los ingresos (X) y estima, con la recta hallada, los gastos fijos de una familia cuyos ingresos son de 1028 € mensuales. ¿Es fiable esta predicción?*

$Y \setminus X$	300-600	600-840	840-1080	1080-1320	1320-1560
36- 72	2				
72-108	1	3	5		
108-144			8	10	
144-180				6	2
180-216					3

Ejercicio 1.7. *Halla y representa las rectas de regresión correspondientes a la distribución estadística:*

X	1	2	3	3	4	4	5	6
Y	5	6	6	7	7	8	8	9
n_{ij}	2	6	7	6	7	4	5	3

Ejercicio 1.8. *Halla la recta de regresión de Y sobre X y analiza la bondad del ajuste realizado:*

X	2	6	10	15
Y	1	2	4	5

Ejercicio 1.9. *Analizados los datos correspondientes a setenta padres de familia en relación a sus retribuciones anuales y su número de hijos, se elaboró la siguiente tabla:*

$Y \setminus X$	0,6 a 1	1 a 1,6	1,6 a 2	2 a 3	3 a 6	6 a 12
0						1
1					3	
2	5	8	17	8		
3	6	18				
4	4					

1. *Suponiendo una relación lineal, determina e interpreta su signo.*

2. *Halla la recta de regresión de Y sobre X y la de X sobre Y.*

3. *¿Qué estimación elegirías para realizar predicciones? ¿Por qué?*

Ejercicio 1.10. *Las ecuaciones:*

$$2x + y + 1 = 0$$
$$5x + 3y + 4 = 0$$

son las rectas de regresión lineal de una distribución estadística bidimensional. Halla el coeficiente de correlación.

Ejercicio 1.11. *Se ha estudiado la correlación existente entre la variable X, que representa los gastos mensuales en publicidad para la promoción de un determinado artículo, y la variable Y, que mide las ventas en el mes siguiente, y se ha obtenido como coeficiente de correlación $r = +0{,}80$. ¿Qué parte del comportamiento de las ventas puede explicarse mediante los gastos en publicidad?*

Ejercicio 1.12. *Analizadas conjuntamente las variables X, que representa el número de alumnos matriculados en la Universidad de Salamanca, e Y, número de visitantes del Museo del Prado, de Madrid, durante cinco años consecutivos, se obtuvo $r^2 = 0{,}87$. ¿Puede asegurarse que el número de visitantes del Museo del Prado depende del número de universitarios de Salamanca?*

Ejercicio 1.13. La tabla que se da a continuación describe ciertas magnitudes relativas a una determinada comunidad autónoma. Los datos de renta y consumo están dados en miles de millones de pesetas.

Año	*Renta*	*Habitantes (en millones)*	*Consumo*
1981	1140	4	798
1982	1360	4,1	986
1983	1800	4,2	1296
1984	2100	4,4	1431
1985	2400	4,6	1658
1986	2730	4,8	2010
1987	3320	5	2257

1. *Designando por Y el consumo y por X la renta per cápita, obtén la función "consumo-renta per cápita" de la forma $Y = aX + b$.*

2. *Estima el consumo para 1988 en el supuesto de que en ese año la renta fuera de 4.008 miles de millones de pesetas, y los habitantes 5,2 millones. Analiza el grado de fiabilidad de la predicción, a través de la medida de bondad de ajuste habitual.*

Ejercicio 1.14. *Decide razonadamente si las ecuaciones*

$$5x + 3y + 8 = 0$$
$$x - 2y - 17 = 0$$

pueden corresponder a las rectas de regresión de una distribución estadística bidimensional.

Ejercicio 1.15. *Se han considerado las puntuaciones obtenidas por un grupo de estudiantes en un test de destreza operativa, realizado a principios de curso, y las calificaciones finales de la asignatura de Matemáticas. Los datos aparecen en la tabla siguiente:*

Calificación final	*Puntuación del test*					
	40-60	*60-80*	*80-100*	*100-120*	*120-140*	*140-160*
3	2	3				
5	2	5				
7			4	6	3	
8			3		1	
9						3

1. *Evalúa el grado de relación lineal que existe entre la destreza operativa y la nota de Matemáticas.*

2. *Asumiendo que la calificación de Matemáticas es una función lineal de la puntuación obtenida en el test, ¿qué efecto produce sobre la nota de Matemáticas un aumento de una unidad en la puntuación del test?*

BIBLIOGRAFÍA COMENTADA

Para una exposición completa y rigurosa de los conceptos tratados en este capítulo, con una notación similar a la aquí empleada, puede verse MARTÍN PLIEGO (MARTÍN PLIEGO, F. Javier. *Introducción a la estadística económica y empresarial (teoría y práctica).* AC. Madrid, 1994), capítulos 7 (Distribuciones bidimensionales) y 9 (Regresión y correlación). El libro contiene, al final de la sección 4 (Análisis estadístico de dos o más variables) una colección de ejercicios (para resolver con calculadora) con solución.

El libro de PEÑA y ROMO (PEÑA, Daniel y ROMO, Juan. *Introducción a la estadística para las ciencias sociales.* McGraw-Hill. Madrid, 1997), capítulos 7, 8 y 9, resulta también muy adecuado. En él abundan los comentarios interpretativos y las explicaciones intuitivas. Por otro lado, contiene datos para realizar aplicaciones con ordenador. Pueden resolverse con un programa estadístico, como STATGRAPHICS, o con una hoja de cálculo, como EXCEL.

Otro libro con explicaciones sencillas y muy intuitivas es el de PÉREZ SUÁREZ (PÉREZ SUÁREZ, Rigoberto. *Análisis de datos económicos I. Métodos descriptivos.* Pirámide. Madrid, 1993). Los capítulos 9 y 11 tratan los temas expuestos y contienen al final una lista de ejercicios con solución. Además, al final del libro hay una serie de problemas propuestos. Con el libro se adjunta un disco con varios ficheros para resolver otros ejercicios con hoja de cálculo.

Los capítulos 5 y 6 del libro de MONTIEL (MONTIEL A. M., RIUS, F. y BARÓN, F. J. *Elementos básicos de estadística económica y empresarial.* Prentice Hall. Madrid, 1997) tratan los temas aquí desarrollados de forma clara y sencilla. En cada capítulo, se presenta una colección de ejercicios resueltos (con calculadora) y un ejemplo práctico resuelto con ordenador con ayuda del programa estadístico SPSS. En el prólogo del libro, y en el desarrollo de cada ejemplo, se dan sencillas instrucciones para el manejo de dicho programa.

2. TEORÍA ELEMENTAL DE LA PROBABILIDAD

1. EXPERIMENTOS ALEATORIOS

Si has jugado al parchís, más de una vez habrás estado detrás de una barrera, deseando que el dueño obtuviera un seis para que la abriera, y poder pasarla. Es éste un claro ejemplo de experimento aleatorio, ya que cada vez que lanzamos un dado no sabemos el resultado que vamos a obtener, por mucho tiempo que llevemos jugando. Los experimentos aleatorios se caracterizan por la imposibilidad de predecir el resultado al repetir el experimento en análogas condiciones.

Se llaman experimentos deterministas aquéllos que producen siempre idéntico resultado al repetirlos en las mismas condiciones, y, por tanto, puede predecirse el resultado de antemano.

La teoría de la probabilidad tiene por objeto el estudio de los experimentos aleatorios.

2. ESPACIO MUESTRAL. SUCESOS. OPERACIONES CON SUCESOS

2.1. Espacio muestral

Al efectuar un experimento aleatorio podemos obtener diversos resultados. Al conjunto formado por todos ellos se le llama espacio muestral, y se le designa con la letra E. Así, al lanzar un dado corriente tenemos:

$$E = \{1,2,3,4,5,6\}$$

Espacio muestral de un experimento aleatorio es el conjunto de todos los resultados posibles de dicho experimento

Ejercicio 2.1. *Escribe el espacio muestral asociado a los siguientes experimentos aleatorios:*

1. *Extraer una carta de una baraja española.*
2. *Lanzar una moneda.*
3. *Lanzar dos dados y observar la suma de puntos obtenida*
4. *Lanzar dos monedas y observar el número de caras.*
5. *Tirar un dado con dos caras rojas, dos verdes y dos azules y una moneda.*

2.2. Sucesos

Cada afirmación referente a los resultados de un experimento aleatorio recibe el nombre de suceso. Por ejemplo, en el experimento del dado serán sucesos "sacar uno", "sacar par", "sacar número primo" o "sacar múltiplo de tres". Observamos que todos ellos son subconjuntos del espacio muestral: {1}, {2,4,6}, {2,3,5} y {3,6}, respectivamente.

Se llama suceso de un experimento aleatorio a cada uno de los subconjuntos del espacio muestral.

Llamaremos suceso elemental al suceso formado por un único resultado del experimento aleatorio.

Decimos que un suceso B está contenido en el suceso A o que implica al suceso A si siempre que se verifica B, se verifica también A. Se escribe $B \subset A$.

Existen dos sucesos especiales en cualquier experimento aleatorio. Es claro que tanto el mismo espacio muestral como el conjunto vacío son subconjuntos de E: $E \subset E$ y $\emptyset \subset E$.

Se llama suceso seguro (E) de un experimento aleatorio aquel que siempre se verifica.

Recibe el nombre de suceso imposible ($\emptyset$) aquel que no se verifica nunca.

Estos dos sucesos, como puede imaginarse, no tienen ningún interés práctico, aunque su uso resulta conveniente cuando se opera con otros sucesos.

Dado un suceso A, se llama suceso contrario de A y se escribe $\overline{A}$, al suceso que se verifica cuando no ocurre A.

Se llama espacio de sucesos y se representa por $P(E)$ al conjunto de todos los sucesos de un experimento aleatorio.

2.3. Operaciones con sucesos

A partir de los sucesos de un experimento aleatorio podemos definir otros sucesos, mediante las operaciones de unión e intersección, que definimos así:

$A \cup B$ es el suceso que ocurre cuando se verifica A o se verifica B (se lee "suceso A o B").

$A \cap B$ es el suceso que ocurre cuando ocurren A y B a la vez (se lee "suceso A y B").

Las propiedades de estas operaciones son las mismas que las de las operaciones con conjuntos:

Conmutativas	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Asociativas	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributivas	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Idempotentes	$A \cup A = A$	$A \cap A = A$
De neutralidad	$A \cup \emptyset = A$	$A \cap E = A$
Absorbentes	$A \cup E = A$	$A \cap \emptyset = \emptyset$
De complemento	$A \cup \overline{A} = E$	$A \cap \overline{A} = \emptyset$
De De Morgan	$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$	$\overline{(A \cap B)} = \overline{A} \cup \overline{B}$
Involutiva	$\overline{\overline{A}} = A$	

Dos sucesos A y B son **incompatibles** si $A \cap B = \emptyset$. En caso contrario se dice que son compatibles.

3. INTRODUCCIÓN A LA PROBABILIDAD

Puede comprobarse experimentalmente que al repetir un gran número de veces en las mismas condiciones un experimento aleatorio, la frecuencia relativa de cada suceso elemental tiende a aproximarse a un cierto valor. Llamaremos a ese valor la **probabilidad del suceso A** (escribimos $p(A)$). Se puede decir que $p(A)$ es el límite de la frecuencia relativa del suceso A cuando el número de pruebas tiende a infinito. Obviamente es imposible conocer el valor de $p(A)$ con esta definición[4].

4. Históricamente, el concepto de probabilidad surge a partir del de frecuencia relativa, dando lugar a la definición frecuentista de probabilidad; esto es, "*el límite de la frecuencia relativa de un suceso cuando el número de tiradas tiende a infinito*". Este hecho empírico, cuya generalización a un experimento cualquiera se conoce como Ley de Regularidad Estadística, afirma que *la frecuencia de un suceso tiende a estabilizarse alrededor de un valor cuando el número de observaciones crece indefinidamente*. Así, la frecuencia relativa del suceso "salir cara al lanzar una moneda" se estabiliza en torno a $\frac{1}{2}$ a medida que el número de tiradas, N, se hace grande, para las monedas nuevas. Esta definición, hoy en desuso, no es operativa ni congruente con el concepto matemático de límite pues se desconoce una expresión general de la frecuencia de un suceso A en función de N, por lo que el límite no es calculable analíticamente y ningún límite puede ser calculado experimentalmente. No obstante, dicha definición consigue dar una idea intuitiva de lo que pretendemos al definir la probabilidad de un suceso. Ésta constituye la definición axiomática de probabilidad, introducida en 1933 por el matemático ruso A. Kolmogorov.

Consideremos, por ejemplo, el experimento de lanzar un dado corriente de seis caras. No parece descabellado suponer que la probabilidad de cada suceso elemental sea 0,1666...= 1/6 (apostaríamos por igual por cualquier número). Esta suposición parece ser confirmada por la experiencia: si realizamos el experimento un gran número de veces, las frecuencias relativas de los sucesos elementales se van aproximando a ese valor. No obstante, desde un punto de vista estrictamente formal, es una decisión arbitraria. En este sentido, podemos asignar a cada suceso elemental de cualquier espacio muestral un número real arbitrario, su probabilidad, con las siguientes limitaciones:

1. Para todo $A \in P(E)$, $0 \leq p(A) \leq 1$; como ocurre con la frecuencia relativa de un suceso, su probabilidad debe ser un número comprendido entre 0 y 1.
2. $\sum_{i=1}^{n} p(A_i) = 1$, siendo $A_1, A_2, \ldots, A_n$ los sucesos elementales, como ocurre con la suma de todas las frecuencias relativas, que es igual a 1.
3. $p(A \cup B) = p(A) + p(B)$, si A y B son sucesos incompatibles.

Desde el punto de vista matemático, cualquier asignación de probabilidades que respete estas condiciones es admisible. Otra cuestión es que refleje una situación real.

Ejemplo. Para participar en las actividades del Instituto se sortean gorras de colores rojo, blanco y verde. Obtendremos una gorra roja cuando saquemos una bola de este color de un bombo que contiene una bola de cada color, y la gorra será blanca o verde si la bola extraída así lo es.

1. Define el espacio muestral.

(Véase Kolmogorov (KOLMOGOROV, A. N. *Foundations of the Theory of Probability*. Chelsea. New York, 1956). La definición frecuentista de probabilidad está propuesta por J. Neyman, cuyos trabajos en este tema pueden consultarse en Pearson y Kendall (PEARSON, E.S. y KENDALL, M.G. *Studies in the History of Statistics and Probability* Ed. Griffin. London, 1970. Pág. 455 y siguientes)

2. Asigna probabilidades a los sucesos elementales.
3. Si supiéramos que de los 200 participantes, 100 han obtenido una gorra roja, 40 una verde y 60 una blanca, ¿qué probabilidades asignarías ahora a esos mismos sucesos?

3.1. Definición de probabilidad

Veamos ahora la definición de probabilidad[5].

Definición. Dado un espacio muestral E, cualquier aplicación $p : P(E) \longrightarrow R$ que verifique:

1. Para todo suceso A de $P(E)$, $0 \leq p(A) \leq 1$.
2. $\sum_{i=1}^{n} p(A_i) = 1$, siendo los A_i los sucesos elementales (y, por tanto, $\bigcup_{i=1}^{n} A_i = E$).
3. Si $A \cap B = \emptyset$, entonces $p(A \cup B) = p(A) + p(B)$

se dice que es una probabilidad definida en E.

Así, es posible definir múltiples probabilidades en un espacio muestral. Cuando definimos una probabilidad en un espacio muestral se dice que tenemos un **espacio probabilístico** o un **espacio de probabilidad**.

Se comprueba que la tercera propiedad puede generalizarse a más de dos sucesos, de la siguiente forma:

3'. Si $A_1, A_2, \ldots, A_n$ son n sucesos incompatibles de dos en dos (esto es, con $A_i \cap A_j = \emptyset$ si $i \neq j$), entonces

$$p(A_1 \cup \ldots \cup A_n) = p(A_1) + \cdots + p(A_n)$$

Además, las propiedades 2 y 3' implican que $p(E) = 1$, como era de esperar.

5. Dicha definición pretende dar una medida relativa y teórica de la ocurrencia de un suceso.

Ejercicio 2.2. *En un yacimiento arqueológico rectangular como el de la figura 2.1, los integrantes de un equipo investigador estiman que la probabilidad de encontrar restos en las zonas A, B y C es 1/5, y la de encontrarlos en D es 2/3. Estudia si los datos son posibles. En caso contrario haz una corrección para que lo sean.*

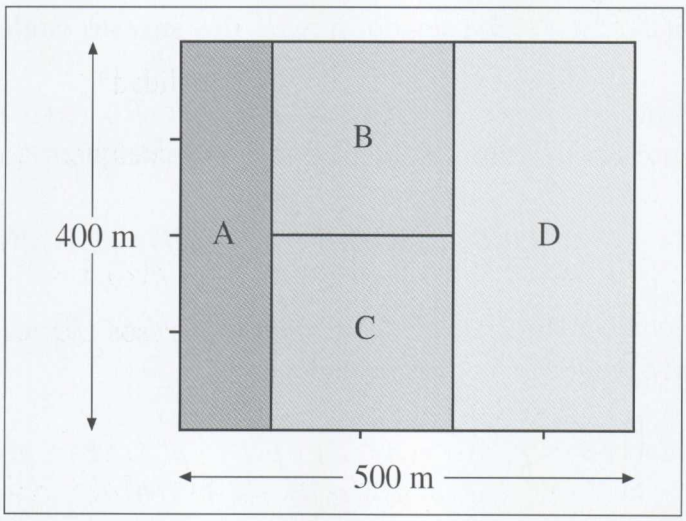

Figura 2.1

3.2. Propiedades de la probabilidad[6]

1. ***Probabilidad del suceso contrario*:** $p(\overline{A}) = 1 - p(A)$.

En efecto, $p(A \cup \overline{A}) = p(A) + p(\overline{A})$, por ser A y $\overline{A}$ incompatibles. Como $A \cup \overline{A} = E$ y $p(E) = 1$ se obtiene el resultado.

2. ***Probabilidad del suceso imposible*:** $p(\emptyset) = 0$.

Es el contrario del suceso seguro.

3. ***Relación entre las probabilidades*: Si** $A \subset B$**, entonces** $p(A) \leq p(B)$.

6. Las propiedades de la probabilidad aquí desarrolladas son paralelas a las de frecuencia relativa. Así, mientras ésta es una medida empírica de la ocurrencia de un suceso, la probabilidad es una medida teórica de la misma, es decir, una medida que trata de evaluar la posibilidad de que ocurra cada suceso antes de realizar el experimento.

Podemos poner $B = A \cup (B \cap \overline{A})$, siendo estos dos últimos sucesos incompatibles.

Como $p(B \cap \overline{A}) \geq 0$, se tiene que $p(B) \geq p(A)$.

4. *Probabilidad de la unión*: $p(A \cup B) = p(B) - p(A \cap B)$, siendo A y B dos sucesos cualesquiera.

Basta con poner $A \cup B = (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (A \cap B)$, siendo estos tres últimos sucesos incompatibles entre sí:

$$p(A \cup B) = p(A \cap \overline{B}) + p(\overline{A} \cap B) + p(A \cap B)$$

Como $p(A \cap \overline{B}) = p(A) - p(A \cap B)$ y $p(\overline{A} \cap B) = p(B) - p(A \cap B)$, sustituyendo:

$$p(A \cup B) = p(A) - p(A \cap B) + p(B) - p(A \cap B) + p(A \cap B) = p(A) + p(B) - p(A \cap B)$$

5. *Regla de Laplace*[7]. En numerosos ejemplos sacados de la vida real, lo más razonable es suponer iguales las probabilidades de todos los sucesos elementales. Son los llamados espacios equiprobables.

Si $E = \{A_1, \ldots, A_n\}$, siendo $p(A_1) = \cdots = p(A_n)$, se tiene que $\forall$ i, $p(A_i) = \frac{1}{n}$. La probabilidad de un suceso $A = A_1 \cup \ldots \cup A_n$

7. La idea clásica de probabilidad que aquí se presenta, introducida por el matemático francés P-S de Laplace (puede consultarse Laplace (LAPLACE, P. S. de. *Théorie analytique des probabilités* (3e édition, 1820), reeditado en 1995). Existe una edición electrónica en http://gallica.bnf.fr/scripts/), no es más que un caso particular de la definición de Kolmogorov planteada en un espacio muestral discreto (donde el conjunto de resultados es finito). Debe reflexionarse sobre el hecho de que la definición propuesta por Laplace no es universal, de hecho es inviable si los resultados del experimento no fueran equiprobables (basta pensar para ello en un dado cargado). Históricamente, esta interpretación *laplaciana* aparece en la literatura estadística a finales del siglo XIX y fue formalizada por R. Von Mises en 1920 (véase Lambalgen (LAMBALGEN, M van. "*Randomness and foundations of probability : von Mises' axiomatisation of random sequences*." En *Statistics, probability and game theory*. IMS Lecture Notes. Hayward, California, 1996. Págs. 347 a 367), sirviendo de motivación trece años después a Kolmogorov para introducir la construcción axiomática de la probabilidad.

será $p(A) = k \times \frac{1}{n} = \frac{k}{n}$, que es la conocida fórmula de Laplace:

$$p(A) = \frac{casos\ \ favorables}{casos\ \ posibles}$$

4. PROBABILIDAD CONDICIONADA. INDEPENDENCIA

Ejercicio 2.3. *De una urna con dos bolas blancas y una negra extraemos una primera bola y, sin devolverla, sacamos una segunda. Supongamos que la primera bola ha sido blanca, ¿cuál será la probabilidad de que también lo fuese la segunda? ¿Y negra en el mismo supuesto?*

Vamos a llamar B_1 = *"Salir blanca la primera bola",*
B_2 = *"Salir blanca la segunda"*
y N_2 = *"Salir negra la segunda bola".*

La probabilidad del suceso B_2 *ha sido condicionada por el hecho de que se ha verificado el suceso* B_1, *ya que con la primera extracción la composición de la urna ha variado. La probabilidad de que la segunda bola sea blanca, sabiendo que lo ha sido la primera, la escribiríamos:* $p(B_2 \mid B_1) = \frac{1}{2}$. *También:* $p(N_2 \mid B_1) = \frac{1}{2}$.

Ejercicio 2.4. *Consideremos el experimento de lanzar un dado, y sean los sucesos A ≡ {Sacar par} y B ≡ {Sacar dos}*

Supongamos que tiramos el dado y nos dicen que se ha verificado A. ¿Cuál será la probabilidad de que se haya verificado B?

Como hemos visto, hay casos en los que obtener información adicional sobre una experiencia modifica la probabilidad asignada a alguno de los sucesos ligados a ella. Conviene aclarar que se trata de

un espacio de probabilidad diferente al original, y que los sucesos no se condicionan, se condicionan las probabilidades[8].

Sea A un suceso de un espacio muestral E de probabilidad no nula ($p(A) \neq 0$). La probabilidad de que haya ocurrido el suceso B, siempre que se haya verificado A en la misma prueba, se llama probabilidad de B condicionada por A}, y se define:

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

Ejercicio 2.5. *En el experimento de sacar una carta de una baraja española sean los sucesos: A≡ {Sacar un rey} y B≡ {Sacar una figura}. Calcular* $p(A)$, $p(A \cap B)$, $p(B \mid A)$ *y* $p(A \mid B)$.

Ejercicio 2.6. *En una reunión de 400 sindicalistas, sabemos que 160 son de U.G.T., 100 de C.C.O.O., 80 de C.S.I.F. y 60 de C.G.T. Se sabe que son mujeres 100 de los delegados de U.G.T., 20 de C.C.O.O., 30 de C.G.T. y 40 de C.S.I.F. Elegido un delegado, resulta ser hombre. Hallar la probabilidad de que pertenezca a cada uno de los sindicatos.*

4.1. Sucesos independientes

Consideremos el experimento de lanzar un dado corriente. Sean los sucesos del espacio muestral A≡ {Sacar par} y B≡ {Sacar número menor que cinco} Es claro que $p(B) = p(B \mid A) = \frac{2}{3}$. Vemos que el hecho de que se haya verificado o no el suceso A no influye en la probabilidad de B.

Introducimos ahora el concepto de independencia[9].

8. De manera similar a lo ya comentado al describir los orígenes del concepto de probabilidad, la probabilidad condicionada tiene sus antecedentes en la frecuencia relativa condicionada, y desde esa perspectiva puede introducirse en el aula.
9. Como ocurre con todos los conceptos del Cálculo de Probabilidades, el de independencia también tiene sus antecedentes en el campo de las frecuencias, pudiendo hablarse, asimismo, de independencia de sucesos en términos de frecuencias condicionadas..

Se dice que un suceso B es independiente de otro suceso A (de probabilidad no nula) si $p(B) = p(B|A)$.

El siguiente resultado es fácil de demostrar:

La independencia es una propiedad recíproca; es decir, siendo A y B sucesos de probabilidad no nula, si B es independiente de A, A lo es de B. Además, en este caso, $p(A \cap B) = p(A) \times p(B)$.

En efecto, por definición se tiene:

$$p(A \cup B) = p(A) \times p(B|A)$$
$$p(A \cup B) = p(B) \times p(A|B)$$

Luego:

$$p(A) \times p(B|A) = p(B) \times p(A|B)$$

Si B es independiente de A, $p(B) = p(B|A)$, y se tiene:

$$p(A) \times p(B) = p(B) \times p(A|B)$$

Así pues, debe ser $p(A) = p(A|B)$.

Si dos sucesos no son independientes diremos que son dependientes.

Ejercicio 2.7. *En un municipio hay tres partidos políticos: Progresista, Liberal y Moderado. Se efectúa un referéndum para decidir si cierto día se declara fiesta local (Sí o No). Los resultados, según el voto de las últimas elecciones municipales, son:*

	Progresista	*Liberal*	*Moderado*	*Abstención*
Sí	*15%*	*25%*	*12%*	*8%*
No	*25%*	*5%*	*8%*	*2%*

Calcular la probabilidad de haber votado a cada uno de los partidos, la probabilidad de haber votado sí y las probabilidades condicionadas p(P|Sí), p(L|Sí), p(M|Sí), p(A|Sí), p(Sí|P), p(S|L) y p(S|A). ¿Haber votado a algún partido es independiente de haber votado sí en el referéndum?

Ejercicio 2.8. *Se lanzan dos dados. Si la suma es seis, hallar la probabilidad de que alguno de ellos sea un dos.*

Ejercicio 2.9. *En una urna hay diez bolas de tres colores, numeradas de la siguiente manera: dos blancas, tres verdes y una roja tienen un uno. Una blanca, dos verdes y una roja tienen un dos.*

1. *Construye un cuadro de doble entrada que represente la situación.*
2. *Calcula p(V), p(1) y p(2|V).*

5. TEOREMAS REFERENTES A LA PROBABILIDAD

Teorema de la probabilidad compuesta (Regla del producto) Sean los sucesos $A_1, A_2, \ldots, A_n$. Se tiene:

$$p(A_1 \cap A_2 \cap \ldots \cap A_n) = p(A_1) \times p(A_2 \mid A_1) \times p(A_3 \mid A_1 \cap A_2) \times \cdots \times p(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

Demostración. Para n =2 se trata de la definición de probabilidad condicionada.

El teorema es inmediato para n =3:

$$p[(A_1 \cap A_2) \cap A_3] = p(A_1 \cap A_2) \times p(A_3 \mid A_1 \cap A_2) = p(A_1) \times p(A_2 \mid A_1) \times p(A_3 \mid A_1 \cap A_2)$$

ya que $p(A_1 \cap A_2) = p(A_1) \times p(A_2 \mid A_1)$.

Para n>3 se generaliza fácilmente.

Ejercicio 2.10. *Sean A y B dos sucesos con probabilidades $p(A)=0{,}5$, $p(B)=0{,}3$ y $p(A\cap B)=0{,}1$.*

Calcular $p(A\mid B), p(A\mid A\cap B), p(A\cap B\mid A\cup B)$ y $p(A\mid A\cup B)$.

Teorema de la probabilidad total.

Vamos a resolver el siguiente problema.

Una fábrica de tornillos tiene tres máquinas, A_1, A_2, A_3. La máquina A_1 produce el 50% de los tornillos, la A_2, el 30% y la A_3, el 20% restante. Se sabe que el 5% de los tornillos producidos por la máquina A_1 son defectuosos, así como el 8% y el 10% de los producidos por A_2 y A_3, respectivamente. De la producción de un día se saca un tornillo al azar. Hallar la probabilidad de que sea defectuoso.

Resolución. Representemos por A_i el suceso de que un tornillo haya sido producido por la máquina A_i y por D que sea defectuoso. En el diagrama de la figura 2.2 se observa que los sucesos A_i son incompatibles dos a dos, lo que lleva a que lo sean los sucesos $A_i \cap D$.

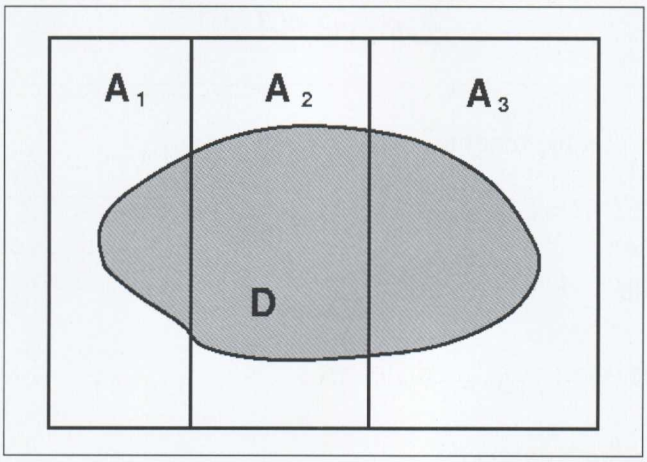

Figura 2.2

Así, se tiene:

$$D=(D\cap A_1)\cup(D\cap A_2)\cup(D\cap A_3)$$

$$p(D)=p(D\cap A_1)+p(D\cap A_2)+p(D\cap A_3)$$

Utilizando la definición de probabilidad condicionada:

$$p(D) = p(D \mid A_1) \times p(A_1) + p(D \mid A_2) \times p(A_2) + p(D \mid A_3) \times p(A_3)$$

Como todos los valores son conocidos:

$$p(D) = 0{,}05 \times 0{,}5 + 0{,}08 \times 0{,}3 + 0{,}1 \times 0{,}2 = 0{,}069$$

Sin mencionarlo, hemos aplicado el **teorema de la probabilidad total**:

Sean los sucesos $A_1, A_2, \ldots, A_n$, tales que cumplan:

$$\bigcup_{i=1}^{n} A_i = E \quad \text{y} \quad \forall \quad i, j \quad A_i \cap A_j = \emptyset .$$

Una familia de sucesos que verifiquen estas condiciones se dice que forman una partición del espacio muestral. Si consideramos cualquier suceso $B \in P(E)$, se tiene:

$$p(B) = p(A_1) \times p(B \mid A_1) + p(A_2) \times p(B \mid A_2) + \cdots + p(A_n) \times p(B \mid A_n) = \sum_{k=1}^{n} p(A_k) \times p(B \mid A_k)$$

En efecto, tenemos

$$B = E \cap B = (A_1 \cup \ldots \cup A_n) \cap B = (A_1 \cap B) \cup \ldots \cup (A_n \cap B)$$

Al ser $A_i \cap B$ y $A_j \cap B$, para todo $i \neq j$, sucesos incompatibles:

$$p(B) = p(A_1 \cap B) + \cdots + p(A_n \cap B)$$

Es claro entonces:

$$p(B) = p(A_1) \times p(B \mid A_1) + p(A_2) \times p(B \mid A_2) + \cdots + p(A_n) \times p(B \mid A_n) = \sum_{k=1}^{n} p(A_k) \times p(B \mid A_k)$$

que es la expresión del teorema.

Ejercicio 2.11. *Los tres delegados de primero de bachillerato de cierto instituto deben elegir un portavoz para quejarse al director de que el profesor de matemáticas "va muy deprisa". Deciden que lo será el que saque el palillo más corto, y no se ponen de acuerdo en quién saca primero, posiblemente por la influencia que tiene este hecho en las probabilidades de que les toque. ¿Qué opinas del asunto?*

Teorema de Bayes[10]**.** Sean los sucesos $A_1, A_2, \ldots, A_n$, tales que formen una partición del espacio muestral E. Si consideramos cualquier suceso $B \in P(E)$, se tiene:

$$p(A_i \mid B) = \frac{p(A_i \cap B)}{p(B)} = \frac{p(A_i) \times p(B \mid A_i)}{\sum_{k=1}^{n} p(A_k) \times p(B \mid A_k)}$$

La demostración es inmediata a partir de la definición de probabilidad condicionada y el teorema anterior.

6. EJERCICIOS PROPUESTOS [11]

Ejercicio 2.12. *Tiramos dos dados y anotamos la suma de los puntos de las caras superiores.*

1. *Escribir los sucesos: A= "Sacar múltiplo de dos"; B= "Sumar más de siete"; C= "Sumar trece".*
2. *Escribir los sucesos contrarios de A y de B.*

10. Debido al reverendo T. Bayes (1702-1761) quien, en obra póstuma (Véase Stigler (STIGLER, S.M. *The History of Statistics. The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. Cambridge, Mass., 1986. Pág. 88) o Pearson y Kendall (PEARSON, E.S. y KENDALL, M.G. *Studies in the History of Statistics and Probability* Ed. Griffin. London, 1970. Pág. 131) Sus trabajos fueron enviados a la Royal Society por su amigo Richard Price en 1764, 3 años después del fallecimiento de Bayes), estudió la probabilidad de las causas ante un efecto observado, sentando las bases de la interpretación subjetiva de la probabilidad y del enfoque bayesiano de la Estadística. En este enfoque, partiendo de una probabilidad (*a priori*), el conocimiento de cierta información nos reasigna una nueva probabilidad (*a posteriori*). Ante una nueva evidencia experimental, corregiremos nuestra creencia primitiva, con lo que la probabilidad se va afinando tras sucesivas incorporaciones de información.

11. Algunos de estos ejercicios, así como otros de diferentes capítulos, están extraídos del libro Fernández-Abascal y otros (FERNÁNDEZ-ABASCAL, H., GUIJARRO, M., ROJO, J. L. Y SANZ, J. A. *Ejercicios de cálculo de probabilidades: resueltos y comentados.* Ariel Matemática. Barcelona, 1995)

Ejercicio 2.13. *Tenemos una baraja española de la que sacamos una carta. Sean los sucesos:*

A= "Obtener oros".
B= "Obtener un número menor que siete".
C= "Obtener una figura".

Escribir los contrarios de cada uno de esos sucesos.

Ejercicio 2.14. *En el experimento que consiste en extraer una carta de una baraja española, sean los sucesos: A="Obtener el as de espadas", B="Obtener un rey" y C="Obtener una carta de oros". Explicar el significado de los siguientes sucesos:*

a) $\overline{A}$	b) $\overline{B}$	c) $\overline{C}$
d) $A \cup B$	e) $B \cap C$	f) $A \cap (B \cap C)$
g) $A \cup (B \cup C)$	h) $\overline{B \cap C}$	i) $A \cup (B \cap C)$

Ejercicio 2.15. *Sean A, B y C tres sucesos del espacio $P(E)$. Se pide expresar en función de ellos y de sus contrarios los sucesos:*

1. *Ocurre A y ocurre B.*
2. *Se realizan A y B, pero no C.*
3. *Ocurre, al menos, uno de los tres.*
4. *No ocurre ninguno de los tres.*

Ejercicio 2.16. *Al tirar un dado, poner un ejemplo de:*

1. *Un suceso seguro.*
2. *El suceso imposible.*
3. *Un suceso A y su contrario.*
4. *Dos sucesos incompatibles.*
5. *Dos sucesos compatibles.*

Ejercicio 2.17. *Tenemos un dado cargado. Estudia las siguientes asignaciones de probabilidades y señala las que cumplen los requisitos de la definición:*

1. p(1)=1/2, p(2)=1/6, p(3)=0, p(4)=1/6, p(5)=0, p(6)=1/6.
2. p(1)=1/3, p(2)=0, p(3)=-1/3, p(4)=2/3, p(5)=1/3, p(6)=0.
3. p(1)=1/3, p(2)=1/3, p(3)=0, p(4)=0, p(5)=1/6, p(6)=1/6, p({1,2})=1/6, p({1,3})=1/3.

Ejercicio 2.18. *En una urna hay un total de 12 bolas entre blancas y negras. Sabemos que la probabilidad de sacar una bola negra es triple que la de sacar una bola blanca. Calcula el número de bolas blancas y negras que hay en la urna.*

Ejercicio 2.19. *Sean* A y B *dos sucesos incompatibles tales que* $p(A)=0{,}3$ y $p(B)=0{,}6$. *Calcúlense las probabilidades de los siguientes sucesos:* $\overline{A}$, $\overline{B}$, $A\cap B$, $A\cup B$, $\overline{A}\cap\overline{B}$ y $\overline{A}\cup\overline{B}$.

Ejercicio 2.20. *Si* A y B *son dos sucesos tales que* $p(A)=0{,}3$, $p(B)=0{,}6$ y $p(A\cap B)=0{,}2$, *hallar las probabilidades de los siguientes sucesos:* $\overline{A}$, $\overline{B}$, $A\cap\overline{B}$, $A\cup B$, $\overline{A}\cup B$, $\overline{A}\cap\overline{B}$ y $\overline{A}\cup\overline{B}$.

Ejercicio 2.21. *Una estadística asegura que el 20% de los habitantes de una ciudad compra habitualmente discos de música clásica, el 30%, discos de rock y el 15%, discos de jazz. Así mismo, el 5% adquiere tanto discos clásicos como de rock, el 7%, de rock y de jazz, el 6%, clásicos y de jazz y el 1%, de los tres tipos. Calcúlese:*

1. *Porcentaje de personas que compran música clásica y no de jazz.*
2. *Porcentaje de individuos que compran discos de rock, o bien de música clásica y de jazz.*
3. *Porcentaje de personas que compran sólo discos de jazz.*

Ejercicio 2.22. *En cierta empresa se ha elaborado un informe sobre las actividades que realizan los empleados en su tiempo libre. Dicho análisis arroja, entre otros, los siguientes resultados: el 30% de los trabajadores practica algún deporte, el 25% dedica varias horas semanales a la lectura y el 10% tiene ambas aficiones.*

1. *Determínese el porcentaje de trabajadores que sólo practica deporte en sus ratos de ocio.*

2. *Calcúlese el porcentaje de empleados que ni lee ni realiza actividades deportivas.*

Ejercicio 2.23. *Lanzamos dos dados y sumamos los puntos obtenidos. Acordamos apostar por parejas. Si la suma es seis u ocho, ganas tú, y si la suma es siete o nueve, gana tu compañero. ¿Te parece un juego equitativo? Justifícalo.*

Lanzamos ahora tres dados y sumamos los puntos obtenidos. Si la apuesta es a favor de la suma nueve o de la suma diez, ¿por cuál apostarías? Halla la probabilidad en cada uno de los casos.

Ejercicio 2.24. *Tres niños y tres niñas se sientan en una fila. Hallar la probabilidad de que las tres niñas se sienten juntas y la de que los niños y las niñas se sienten alternados.*

Ejercicio 2.25. *Se carga un dado de modo que los números pares tienen doble probabilidad de salir que los impares. Hallar la probabilidad de que:*

1. *Aparezca un número par.*
2. *Aparezca un número primo.*
3. *Aparezca un número impar.*
4. *Aparezca un número primo impar.*

Ejercicio 2.26. *De 120 estudiantes, 60 estudian inglés, 50 estudian francés, y 20 estudian francés e inglés. Si se escoge un estudiante al azar, hallar la probabilidad de que el estudiante:*

1. *Estudie francés e inglés.*
2. *No estudie ni francés ni inglés.*
3. *Estudie francés pero no inglés.*

Ejercicio 2.27. *Se selecciona una carta al azar entre 50 cartas numeradas de uno a cincuenta. Hallar la probabilidad de que el número de la carta sea:*

1. *Divisible por 5.*
2. *Primo.*
3. *Termine en 2.*

Ejercicio 2.28. *Hallar las probabilidades de que al lanzar al aire tres veces un dado la suma de los puntos sea:*

1. *Múltiplo de* 5.
2. *Mayor que* 4.

Ejercicio 2.29. *Un dado en forma de dodecaedro regular está construido de tal forma que la probabilidad de obtener un número determinado es proporcional a dicho número (las caras van marcadas del 1 al 12) Calcular la probabilidad de obtener:*

1. *Un número par.*
2. *Un número múltiplo de* 4.
3. *Un número mayor o igual que* 10.

Ejercicio 2.30. *En una clase, el 25% de los alumnos ha suspendido las matemáticas, el 15%, la química y el 10%, las matemáticas y la química. Escogido un estudiante al azar, hallar la probabilidad de que:*

1. *Haya suspendido la química si ha suspendido las matemáticas.*
2. *Haya suspendido las matemáticas si ha suspendido la química.*
3. *Haya suspendido, al menos, una de las dos asignaturas.*

Ejercicio 2.31. *Un opositor ha preparado 40 temas de los 100 que tiene el temario. El examen consiste en desarrollar tres de los temas elegidos al azar. ¿Qué probabilidad tiene de que le caigan tres de los que se sabe?*

Ejercicio 2.32. *En una baraja española de 40 cartas se toman dos al azar con reemplazamiento.*

1. *Calcular la probabilidad de que las dos sean ases.*
2. *Los sucesos "La primera carta es un as" y "La segunda carta es un as", ¿son independientes?*
3. *Contesta a las mismas preguntas sin reemplazamiento.*

BIBLIOGRAFÍA COMENTADA

No queremos dejar de señalar una interesante introducción histórica, que recopila y comenta los primeros textos publicados sobre Teoría de la Probabilidad. Se trata del libro de De Mora Charles (MORA CHARLES, Marisol de. *Los inicios de la Teoría de la Probabilidad. Siglos XVI y XVII.* Servicio editorial UPV. Bilbao, 1989) que, a buen seguro, constituye una fuente inagotable de sugerencias sobre los juegos de azar, la naturaleza estadística de los milagros y de la existencia de Dios, el tratamiento estadístico de los censos, etc.

El manual de Fdez-Abascal y otros (FERNÁNDEZ-ABASCAL, H., GUIJARRO, M., ROJO, J. L. Y SANZ, J. A. *Cálculo de Probabilidades y Estadística.* Ariel Economía. Barcelona, 1994) constituye una frecuente y socorrida fuente de referencia; así, en sus capítulos 2 y 3 se detallan hasta extremos no tratados habitualmente en otros manuales, conceptos difusos o *difíciles* de entender y sobre todo de explicar. Por ejemplo, es excelente la exposición dedicada a la probabilidad en el espacio producto, de la que otros manuales al uso apenas dedican una referencia.

Un libro clásico en estos temas, en inglés, es Gnedenko (GNEDENKO, B. V. *The Theory of Probability.* MIR. Moscú, 1978), así como Cramér (CRAMÉR, H. *Elementos de la Teoría de probabilidades y algunas de sus aplicaciones.* Aguilar. Madrid, 1977). Anderson (ANDERSON, I. *Introducción a la Combinatoria.* Vicens Vives. Barcelona, 1993) o Ardanuy y Soldevilla (ARDANUY, R. Y SOLDEVILLA, M. M. *Estadística Básica*, Hespérides. Salamanca, 1992), que pueden ayudar a la comprensión de la Combinatoria, no muy desarrollada en el programa. No obstante, es Feller (FELLER, W. *Introducción a la teoría de Probabilidades y sus aplicaciones.* Vol. I y II. Limusa. México, 1975), en su capítulo II, el manual clásico por excelencia en cuanto al tratamiento del Cálculo de probabilidades desde una perspectiva combinatoria.

Finalmente, entre los manuales de estadística general con apartados dedicados a la probabilidad destacan el capítulo 1 de Walpole y Myers (WALPOLE, R. E. Y MYERS, R. H. *Probabilidad y Estadística*. McGraw Hill. 4ª. ed. México, 1991) y Barbancho (GARCÍA BARBANCHO, A. *Estadística teórica básica: Probabilidad y modelos probabilísticos*. Ariel Economía. Barcelona, 1992). Durá y López (DURÁ. J. M. Y LÓPEZ, J. M. *Fundamentos de Estadística: Estadística descriptiva y modelos probabilísticos para la inferencia*. Ariel Economía. Barcelona, 1988), capítulo 4; Fernández y Fuentes (FERNÁNDEZ, C. Y FUENTES, F. *Curso de Estadística descriptiva. Teoría y práctica*. Ariel Economía. Barcelona, 1995), capítulo 5; Fdez de Trocóniz (FERNÁNDEZ DE TROCÓNIZ, A. *Introducción a las teorías de las Probabilidades. Estadística clásica y Estadística bayesiana*. Autoeditado. Bilbao, 1980), capítulos 1 a 4; Escuder y Murgui (ESCUDER, R. Y MURGUI, J. S. *Estadística aplicada. Economía y Ciencias sociales*. Tirant lo blanch. Valencia, 1995), capítulo 13. Merece señalarse especialmente Casas y Santos (CASAS, J. Y SANTOS, J. *Introducción a la Estadística para economía y administración de empresas*. Centro de Estudios Ramón Areces. Madrid, 1995), en sus capítulos 1, 7 y 8. Nortes (NORTES, A. *Estadística teórica y aplicada*. DM y PPU. Murcia y Barcelona, 1991), capítulo 6, plantea muchos ejercicios y Canavos (CANAVOS, G. C. *Probabilidad y Estadística: Aplicaciones y Métodos*. McGraw Hill. México, 1992), capítulo 2, desarrolla ejemplos reales para introducir cada concepto. Cuadras (CUADRAS, C. M., ECHEVARRÍA, B., MATEO, J. Y SÁNCHEZ, P. *Fundamentos de estadística: Aplicación a las Ciencias humanas*. P.P.U. Barcelona, 1984), dedica el capítulo 1 a la combinatoria, y el 3 y 5, con detalle, al espacio producto. DeGroot (DEGROOT, M. H. *Probabilidad y Estadística*. Addison-Wesley Iberoamericana. México, 1988), capítulos 1 y 2, realiza sugerentes comentarios acerca del uso engañoso de la Estadística. De sus usos y abusos también trata el capítulo 0 de Chao (CHAO, L. L. *Estadística para las ciencias administrativas*. McGraw Hill. 3a. ed. Bogotá, 1993).

Un libro *distinto* en cuanto a su concepción y metodología, pero muy educativo para el docente es Llopis (LLOPIS PÉREZ, J. *La estadística: una orquesta hecha instrumento*. Ariel Ciencia. Barcelona,

1996), que propone un discurso constructivo de la probabilidad y del uso y asignación de distribuciones a las variables aleatorias.

Martín-Pliego y Ruiz Maya (MARTÍN PLIEGO, F. J. Y RUIZ MAYA, L. *Estadística I: Probabilidad*, AC. Madrid, 1995), capítulos 0 y 1, ponen en relación las distintas concepciones de la probabilidad, discusión que también aborda López Cachero (LÓPEZ CACHERO, M. *Fundamentos y métodos de estadística*. Pirámide. Madrid, 1989) en su capítulo 14. Señalemos también los sugerentes comentarios acerca de *posibilidad* y *probabilidad* realizados en Freund y Simon (FREUND, J. E. Y SIMON, G. A. *Estadística Elemental*. 8a. ed. Prentice-Hall. México, 1994). Debemos citar, asimismo, el trabajo de Kendall (KENDALL, M. G. *Enciclopedia Internacional de las Ciencias Sociales*. Vol 4, Aguilar. Madrid, 1979. Pág. 404).

Rohatgi (ROHATGI, V. K. *An Introduction to Probability Theory and Mathematical Statistics*. Wiley. New York, 1977) y (ROHATGI, V. K. *Statistical Inference*. Wiley. New York, 1994), con uno de los tratamientos y desarrollos más rigurosos, matemáticamente hablando, aborda todo el planteamiento inferencial, así como la formalización del Cálculo de Probabilidades.

En cuanto a libros de problemas, cabe citar, entre otros muchos a Fdez-Abascal y otros (FERNÁNDEZ-ABASCAL, H., GUIJARRO, M., ROJO, J. L. Y SANZ, J. A. *Ejercicios de cálculo de probabilidades: resueltos y comentados*. Ariel Matemática. Barcelona, 1995), capítulos 1 y 2.

3. VARIABLES ALEATORIAS. DISTRIBUCIONES DE PROBABILIDAD

1. VARIABLES ALEATORIAS

En los ejemplos que a continuación siguen, aparecen espacios muestrales. Cada uno de ellos se definía como un conjunto que contiene los resultados posibles de un experimento. Éstos podían ser de dos tipos:

- de tipo **cuantitativo.** Por ejemplo, resultados obtenidos al lanzar un dado, al tomar la estatura a los alumnos, resultados del juego de la ruleta, el número premiado en la lotería del último sábado, etc.
- de tipo **cualitativo**. Por ejemplo, sexo de un recién nacido, resultado al lanzar una moneda, barrio de procedencia de tus compañeros de clase, etc.

Tanto en unos casos como en otros, interesa poner en correspondencia los resultados posibles del experimento aleatorio con números reales, definiendo de este modo lo que se denomina variable aleatoria; se trata, por tanto, de una característica numérica a la que, bajo ciertas condiciones, asociaremos una probabilidad.

Ejemplo 1. Efectuamos el experimento aleatorio consistente en lanzar un dado y observar el resultado. El conjunto de los resultados posibles del mismo, el espacio muestral, es:

$$E = \{1, 2, 3, 4, 5, 6\}$$

siendo los resultados de tipo cuantitativo.

Sucede que dos amigos deciden jugar a par o impar, apuntándose 1 punto si obtienen par, y restándoselo si obtienen impar.

Así, X será una aplicación del espacio muestral E en el conjunto formado por dos elementos $\{-1,1\}$, de modo que los números pares $\{2\}$, $\{4\}$ y $\{6\}$ se aplicarán en 1 y los resultados impares $\{1\}$, $\{3\}$ y $\{5\}$ en -1, de la siguiente forma:

$$X(1) = X(3) = X(5) = -1$$

$$X(2) = X(4) = X(6) = 1$$

Diremos entonces que X es una variable aleatoria, pues lleva (aplica) resultados muestrales en números reales.

Ahora, otros dos amigos plantean un juego similar, de modo que ganan un punto si se obtiene un resultado entre $\{1,2,3\}$ y pierden un punto si es de los tres últimos. En este caso, definimos la variable aleatoria Y:

$$Y(1) = Y(2) = Y(3) = -1$$

$$Y(4) = Y(5) = Y(6) = 1$$

que lo es del mismo espacio muestral que el anterior, el de los resultados de un dado.

Podemos, en consecuencia, disponer de muchas variables aleatorias para un mismo espacio muestral. ¿Se te ocurre alguna otra variable aleatoria asociada a ese espacio muestral?

Ejemplo 2. El experimento en esta ocasión consiste en lanzar una moneda dos veces y observar el resultado. El espacio muestral asociado, espacio de los posibles resultados, está formado por 4 elementos:

$$E = \{(c,c),(c,z),(z,c),(z,z)\}$$

donde designamos por c el resultado obtener cara y por z, obtener cruz. Admitamos que la moneda es perfecta, con lo que podemos suponer que todos los resultados son igualmente probables, es decir:

$$p[(c,c)]=p[(c,z)]=p[(z,c)]=p[(z,z)]=\frac{1}{4}$$

Ahora bien, nos interesa más estudiar el *número de caras obtenidas*, característica numérica de los resultados del experimento que designaremos por X, que el estudiar qué resultado muestral ha salido.

X es una aplicación que a cada resultado del experimento le hace corresponder un número real, en este caso, el número de caras que se observan al lanzar una moneda dos veces. La figura 3.1 recoge esta situación.

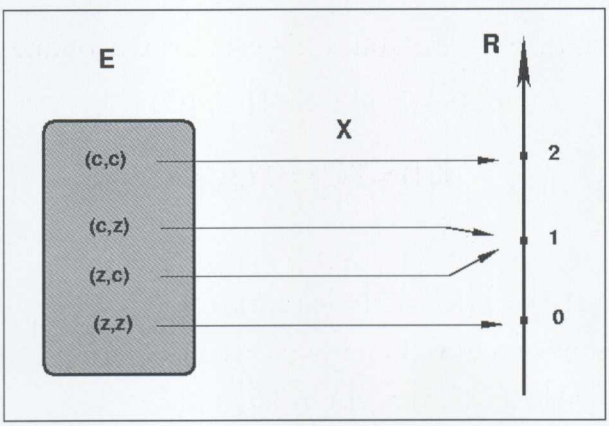

Figura 3.1

De este modo, la probabilidad de obtener una cara será la probabilidad del conjunto de resultados del experimento en los que se observa una sola cara, a saber:

$$p[X=1]=p[(c,z),(z,c)]=p[(c,z)]+p[(z,c)]=\frac{1}{4}+\frac{1}{4}=\frac{1}{2}$$

Pudiéndose obtener otras probabilidades directamente, como la de obtener al menos una cara

$$p[X \geq 1] = p[(c,z),(z,c),(c,c)] = \frac{3}{4}$$

o la de obtener estrictamente menos de dos caras

$$p[X < 2] = p[(c,z),(z,c),(z,z)] = \frac{3}{4}$$

La función de distribución[12] de una variable aleatoria, X, es una función numérica, es decir, lleva números en números y va a permitir calcular probabilidades para la variable aleatoria. Juega un papel similar al de la tabla de frecuencias acumuladas de una variable estadística, cuando utilizábamos las distribuciones de frecuencias. Se denota como $F_X(x)$ y se define como la probabilidad hasta el punto x.

Calculemos $F_X(x)$ para diferentes números reales x, en este ejemplo; así:

$$F_X(-1) = p[X \leq -1] = p(\emptyset) = 0$$

$$F_X(0{,}5) = p[X \leq 0{,}5] = p[X = 0] = p[(z,z)] = \frac{1}{4}$$

$$F_X(1{,}5) = p[X \leq 1{,}5] = p[(z,z),(c,z),(z,c)] = \frac{3}{4} \quad \text{ó}$$

$$F_X(2{,}5) = p[X \leq 2{,}5] = p(E) = 1$$

Formalmente, hay que calcular la función de distribución para cada número real x, igual que para los cuatro anteriores; así, si x es cualquier número entre 0 y 1, es decir, en el intervalo [0, 1), la función de distribución en él vale siempre 1/4, pudiéndose obtener, con un argumento similar, el resto de sus valores:

12. La función de distribución de una variable aleatoria es la función que describe la acumulación de probabilidad asociada a la variable a lo largo de la recta real. Así, tramos con fuerte crecimiento indican zonas de valores de la variable que aportan gran cantidad de probabilidad, mientras que tramos planos evidencian zonas en los que no hay aporte alguno.

$$F_X(x)=\begin{cases} 0 & si \quad x<0 \\ 1/4 & si \quad 0\le x<1 \\ 3/4 & si \quad 1\le x<2 \\ 1 & si \quad x\ge 2 \end{cases}$$

La representación gráfica de esa función viene dada por la siguiente figura.

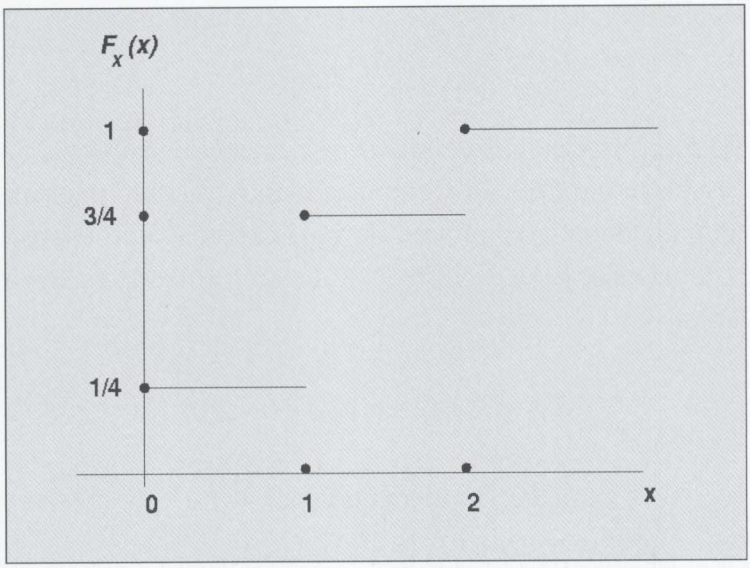

Figura 3.2

Nota: Las desigualdades "menor o igual" ($\le$) y "menor estricto" ($<$) no están puestas al azar. Si cambiamos alguna de ellas, cambia la función. En concreto, puedes comprobar que para x=2, $F_X(2)$ no vale 3/4 sino 1.

Ejemplo 3. Se plantea estudiar la distribución por sexos de una familia de tres hijos. En este caso, el espacio muestral es

$$E=\{(h,h,h),(h,h,m),(h,m,h),(m,h,h),(h,m,m),(m,h,m),(m,m,h),(m,m,m)\}$$

donde con h expresamos "hombre" y con m, "mujer", resultados de tipo cualitativo. Sea X la variable aleatoria número de hombres e Y, número de mujeres. Trata de hacer un diagrama como el anterior y que recoja el valor numérico de cada uno de los resultados posibles del espacio muestral para cada una de las variables.

Ejemplo 4. El experimento propuesto en esta ocasión consiste en elegir a una persona de tu aula e interrogarla sobre su peso y su altura. Así, el espacio muestral estará formado por todos los alumnos del aula

$$E = \{Alumno_1, Alumno_2, \cdots\}$$

Los resultados en este caso son también cualitativos. Ejemplos de variables aleatorias son el peso del alumno elegido, X, y la altura del mismo, Y.

Ejemplo 5. Disponemos de una urna que contiene 4 bolas negras y 2 blancas y extraemos 3 sin reemplazamiento, estando interesados en la variable aleatoria X, número de bolas blancas que aparecen tras las 3 extracciones. El espacio muestral está formado por los resultados cualitativos siguientes:

$$E = \{(3n), (2n, \text{y } 1b), (1n \text{ y } 2b)\}$$

donde, como es evidente, n denota obtener negra y b, obtener blanca y el dígito previo, el número de éstas.

La figura 3.3 esquematiza esta situación de forma análoga a como lo estamos haciendo.

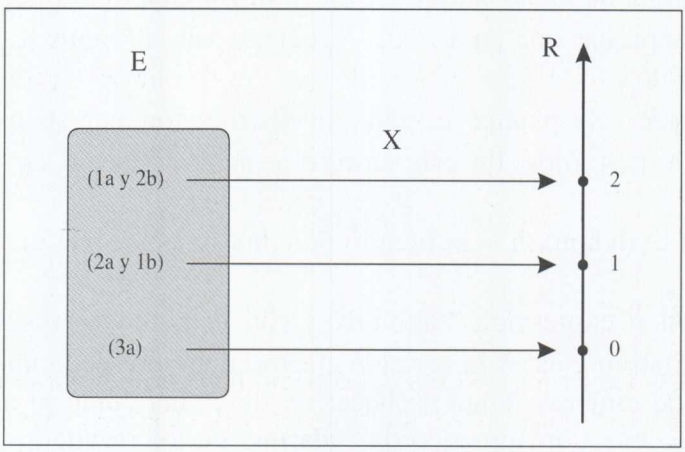

Figura 3.3

Intenta dibujar la misma situación para la variable aleatoria Y, número de bolas negras extraídas.

¿Te atreverías a aventurar qué sucedería si tras una extracción de la urna, devolviéramos la bola?, es decir, ¿si las extracciones son con reemplazamiento? Escribe el espacio muestral y dibuja las dos aplicaciones (las variables aleatorias) anteriores X e Y.

Ejemplo 6. Similar al ejemplo 1. Tomamos ahora dos dados, estando interesados en la suma de los dos resultados. En este caso, el espacio muestral de interés (los resultados posibles del experimento) resulta ser

$$E = \{2,3,4,5,6,7,8,9,10,11,12\}$$

y no el que en un principio definiríamos como

$$E = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & \cdots, & (1,6), \\ (2,1), & (2,2), & (2,3), & \cdots, & (2,6), \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (6,1), & (6,2), & (6,3), & \cdots & (6,6) \end{Bmatrix}$$

Ahora bien, el cálculo de las probabilidades de los once resultados de E no se simplifica (de hecho, deberíamos recurrir, explícita o implícitamente, a las probabilidades definidas sobre E). Además, los resultados obtenidos no nos servirían para estudiar otras características (salvo, obviamente, para las que se obtengan a partir de la suma)

Cada uno de los sucesos elementales de E, tiene probabilidad 1/36 ya que

$$p[(i,j)] = p(i) \times p(j) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36},$$

siendo, además, finito el espacio muestral E.

¿Cuáles de los sucesos del espacio muestral E garantizan una suma de resultados igual a 2?; únicamente el (1,1), esto se expresa como

$$[X = 2] = \{(1,1)\},$$

pudiéndose afirmar que la probabilidad de que X valga 2 es la probabilidad de que al lanzar dos dados, obtengamos dos unos, es decir,

$$p[X=2]=p(\{(1,1)\})=p(\{1\})\times p(\{1\})=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36},$$

En este contexto, si en el juego ganamos cuando la suma de los resultados de ambos dados es menor o igual que 5, lo haremos cuando ocurra alguno de los siguientes

$$[X\le 5]=\{(1,1),(1,2),(1,3),(1,4),(2,1),(2,2),(2,3),(3,1),(3,2),(4,1)\}$$

teniendo una probabilidad de

$$p[X\le 5]=10\times\frac{1}{36}=0{,}2778.$$

En idéntico sentido, los resultados que llevan a una suma mayor o igual que diez son:

$$[X\ge 10]=\{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$$

siendo su probabilidad

$$p[X\ge 10]=6\times\frac{1}{36}=\frac{1}{6}.$$

Ejemplo 7. Lanzamos sucesivamente una moneda hasta que obtengamos la primera cara y llamamos X a la variable aleatoria "número de tiradas hasta obtener la primera cara". Los resultados posibles son

$$E=\left\{c,(z,c),(z,z,c),(z,z,z,c),\ldots,(z,\overbrace{\ldots}^{k)},z,c),\ldots\right\}$$

donde c denota que ha salido cara y z que ha salido cruz, y X valdrá 1, 2,..., k,... según qué resultado muestral hayamos obtenido.

En cualquiera de los siete ejemplos anteriores, disponemos de un espacio muestral, E, asociado a un cierto experimento aleatorio y una(s) correspondencia(s) entre los elementos de E y los números reales R. A esa correspondencia X, se la denomina **variable aleatoria**[13] del espacio muestral E.

Si X toma, o puede tomar, un número finito o infinito numerable de valores, se dice que X es una variable aleatoria discreta. Por ejemplo, son discretas, el número de hijos, el número de accidentes laborales, el número de ingresos en un hospital, etc.

Por contra, si la característica numérica a estudiar puede tomar una cantidad infinita no numerable de valores (toda la recta real o una parte de ella), aparece el concepto de variable aleatoria continua. Por ejemplo, altura, peso, renta, tiempo, etc.

Debemos hacer notar que la diferencia se encuentra en la precisión de las mediciones. Al decir que una persona mide 180 cm, entenderemos que su estatura es un valor entre 179,5 y 180,5. Su valor exacto es imposible de conocer y vendrá condicionado por el aparato de medida utilizado, con el que siempre se comete un error aunque sea muy pequeño.

Por establecer una comparación, piensa en las variables estadísticas agrupadas en intervalos. En ellas, al pasar a intervalos de menor longitud, va consiguiéndose el aspecto que puede apreciarse en las gráficas que acompañan al ejemplo 8.

Aunque el aspecto es similar, conseguimos mayor información estrechando el intervalo de modo que, siguiendo el proceso,

13. Las variables aleatorias no son sino aplicaciones, no todas, del espacio muestral en la recta real. Por tanto, son asignaciones de valores que se realizan en un espacio muestral. Por motivos históricos se usa el nombre de *variables aleatorias* en vez de *funciones* que es lo que son realmente, y se describen con las últimas letras del abecedario, U, V, W, X, Y o Z, en mayúsculas. El calificativo *aleatorias* que acompaña su denominación, sirve para diferenciarlas de su antecedente histórico, las variables estadísticas. Éstas, recogían información numérica observada de los individuos de una población, y en este sentido contabilizaban los valores observados en términos de frecuencias relativas. Por el contrario, los posibles valores de una variable aleatoria vendrán contabilizados en términos de probabilidad.

llegaríamos a una curva como representante de la distribución de la variable que estemos estudiando.

Esta curva se conoce con el nombre de función de densidad y se representa como $f(x)$ debiendo cumplir[14]:

1. Ser positiva $f(x) \geq 0$
2. El área comprendida entre la curva y el eje de abscisas es 1. Esto es, la integral en todo el campo donde la función no sea nula será 1:
   $\int_a^b f(x)dx = 1$ siendo (a,b) el intervalo donde $f(x)$ no es nula.

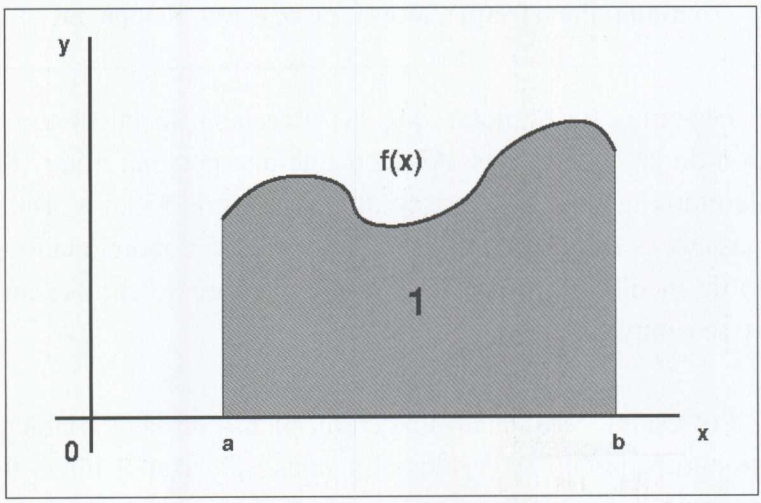

Figura 3.4

Veamos un ejemplo.

Ejemplo 8. Se mide la altura X de los 1000 alumnos del Instituto. Agrupando los datos en intervalos de anchura 10 cm se obtienen los siguientes resultados.

14. Debe tenerse precaución en este terreno. La función de densidad en un punto x no es la probabilidad en dicho punto, es la probabilidad por unidad de longitud en un intervalo infinitesimal alrededor de x. Es decir, expresa la densidad de probabilidad en torno al punto. Analíticamente lo expresamos:

$$f(x) = \lim_{h \to 0} \frac{p\left(x - \frac{h}{2} \leq X \leq x + \frac{h}{2}\right)}{h}$$

Altura (cm)	n_i	f_i
135-145	7	0,007
145-155	60	0,060
155-165	242	0,242
165-175	383	0,383
175-185	241	0,241
185-195	61	0,061
195-205	6	0,006

Correspondiendo el histograma siguiente a su distribución.

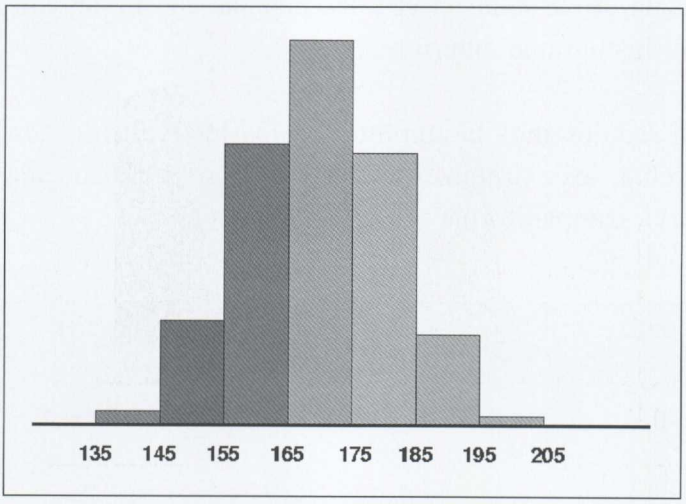

Figura 3.5

Con este histograma podemos responder a cuestiones del tipo, ¿con qué frecuencia se presentan los valores de la variable comprendidos entre 150 y 170?, ¿y los menores de 195? Asimismo, podemos obtener el número de alumnos en cada una de las dos situaciones anteriores.

Afinemos el proceso de medición de alturas, para ello consideramos intervalos de longitud 5 cm, resultando el siguiente histograma.

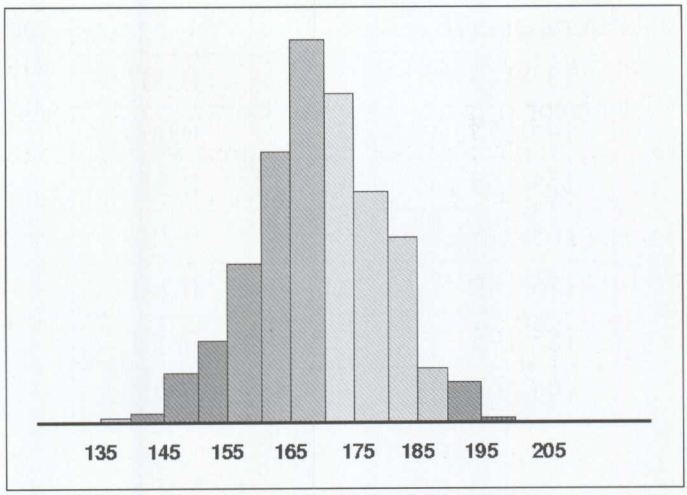

Figura 3.6

Trata de calcular en él a las mismas cuestiones que planteamos en el histograma anterior.

Si redujésemos la amplitud de los intervalos, al tener barras más estrechas, los extremos superiores se aproximarían cada vez más a una curva campaniforme como la siguiente.

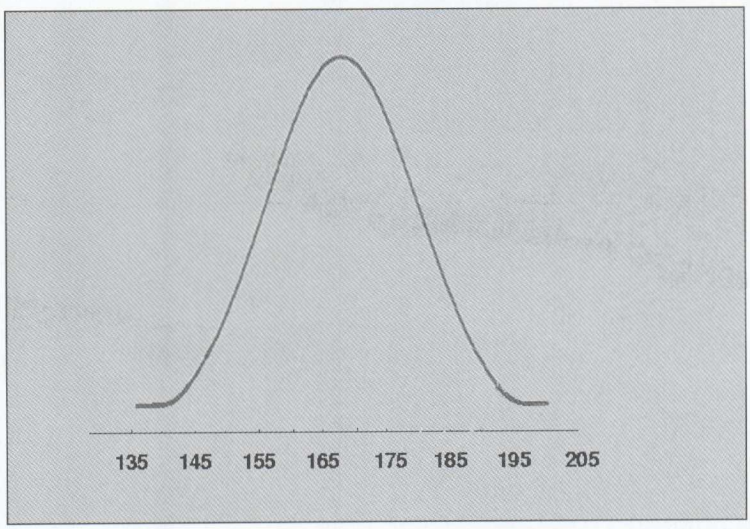

Figura 3.7

Este tipo de curvas de cuyo comportamiento hablaremos posteriormente tiene una gran importancia.

Señalemos que la función de densidad no nos permite calcular probabilidades "puntuales". Por ejemplo, $f(180)$ es un número real mayor o igual que cero, pero no es la probabilidad de que la variable tome el valor 180. Como ya señalábamos antes, aunque $[X=180]$ sea un valor de la variable, ¿con qué precisión mediríamos para poder hablar de $p[X=180]$? Sólo tendría sentido empírico el que nos refiramos a la probabilidad de que la variable esté en cierto intervalo alrededor de 180. Por tanto, sólo tendrán sentido $p[X<a], p[X\leq a], p[a<X\leq b]$ o expresiones similares.

Los ejemplos previos (salvo el 4º y el 8º, que son de una variable aleatoria continua), son todos de variables aleatorias discretas, siendo finito el conjunto de los valores posibles (salvo en el 7º, que toma una cantidad infinita numerable de valores).

Por lo general, trabajaremos con variables aleatorias discretas y finitas.

Retomemos el ejemplo 3. En él estábamos interesados en la variable aleatoria X, número de hombres en una familia de 3 hijos. Del espacio muestral

$$E=\{(h,h,h),(h,h,m),(h,m,h),(m,h,h),(h,m,m),(m,h,m),(m,m,h),(m,m,m)\}$$

deducimos que X puede tomar cualquiera de los siguientes valores, $\{0,1,2,3\}$.

$$[X=0]=\{(m,m,m)\}\mapsto \qquad p[X=0]=p(m)\times p(m)\times p(m)=\frac{1}{2^3}=\frac{1}{8}$$

$$[X=1]=\{(m,m,h),(m,h,m),(h,m,m)\}\mapsto \qquad p[X=1]=p(m,m,h)+p(m,h,m)+p(h,m,m)=3\times\frac{1}{2^3}=\frac{3}{8}$$

$$[X=2]=\{(m,h,h),(h,h,m),(h,m,h)\}\mapsto \qquad p[X=2]=p(m,h,h)+p(h,h,m)+p(h,m,h)=3\times\frac{1}{2^3}=\frac{3}{8}$$

$$[X=3]=\{(h,h,h)\}\mapsto \qquad p[X=3]=p(h)\times p(h)\times p(h)=\frac{1}{2^3}=\frac{1}{8}$$

Con esto, podemos resumir la historia de la variable en la siguiente tabla

x_i	p_i
0	1/8
1	3/8
2	3/8
3	1/8
	1

Sea X una variable aleatoria discreta y E un espacio muestral formado por $\{x_1,\ldots,x_n\}$, conjunto finito de valores que puede tomar la variable. Se tiene

$$E=[X=x_1]\cup[X=x_2]\cup\cdots\cup[X=x_n]$$

es decir, parten el espacio muestral en sus sucesos elementales.

A la probabilidad del resultado x_i se la denota por $p_i=p[X=x_i]$.

Habitualmente, se expresa la variable acompañada de sus probabilidades. A esto se le llama la función de probabilidad para una variable aleatoria discreta y finita.

Obsérvese que $p_1+\cdots+p_n=1$, ya que es la suma de las probabilidades de todos los sucesos posibles.

Nótese la absoluta similitud entre esta tabla y la que existía con una variable estadística en donde aparecían las frecuencias relativas.

Si llamamos $F_i=p_1+\cdots+p_i$, es decir, la probabilidad hasta el punto x_i, sucede como ocurría en la Estadística Descriptiva con las frecuencias acumuladas, dando lugar a las probabilidades acumuladas o distribución de probabilidad. En general se expresa como sigue

x_i	p_i	F_i
x_1	p_1	p_1
x_2	p_2	$p_1 + p_2$
$\vdots$	$\vdots$	$\vdots$
x_i	p_i	$p_1 + p_2 + \cdots + p_i$
$\vdots$	$\vdots$	$\vdots$
x_n	p_n	1

donde, por ejemplo:

$$F_3 = p_1 + p_2 + p_3 = p[X = x_1] + p[X = x_2] + p[X = x_3] = p[X \leq x_3]$$

Planteemos el siguiente ejemplo.

Ejemplo 9. Una caja contiene tres bolas: una blanca, otra roja y la tercera negra. Consideremos la variable aleatoria X, *número total de bolas blancas* obtenidas en dos extracciones con reemplazamiento. El espacio muestral asociado al experimento es:

$$E = \{(b,b),(b,r),(b,n),(r,b),(r,r),(r,n),(n,b),(n,r),(n,n)\}$$

donde *b, r, n* indican, respectivamente, sacar bola blanca, roja y negra. Razonablemente asignaremos a cada uno de esos sucesos la misma probabilidad, igual a 1/9. Entonces, la variable aleatoria X es una aplicación del espacio muestral E en R tal que

$$X[(b,b)] = 2$$
$$X[(b,r)] = X[(b,n)] = X[(r,b)] = X[(n,b)] = 1 \quad \text{y}$$
$$X[(r,r)] = X[(r,n)] = X[(n,r)] = X[(n,n)] = 0$$

Por tanto, el conjunto de valores de la variable X es:

$$\{0,1,2\}$$

La probabilidad de que el número de bolas blancas sea, como mucho, de 1, es la probabilidad del conjunto $[X \leq 1]$, y es

$$p[X \leq 1] = p[(b,r),(b,n),(r,b),(r,r),(r,n),(n,b),(n,r),(n,n)] = \frac{8}{9}.$$

Adviértase que dos o más variables pueden tener la misma distribución. Por ejemplo, X, *número de caras obtenidas en el lanzamiento de una moneda*, e Y, *número de resultados pares obtenidos al tirar un dado*, son variables distintas, pero si se supone que tanto la moneda como el dado están sin cargar, es decir, sin trampa o truco, ambas tienen la misma distribución. Así, ambas variables las expresamos como

$$X: \begin{vmatrix} 1 & si & cara \\ 0 & si & cruz \end{vmatrix}$$

e

$$Y: \begin{vmatrix} 1 & si & par \\ 0 & si & impar \end{vmatrix}$$

pero

$$p[X = 1] = p[Y = 1] = \frac{1}{2}$$

y

$$p[X = 0] = p[Y = 0] = \frac{1}{2}$$

Por tanto,

$$F_X(x) = \begin{cases} 0 & si & x < 0 \\ 1/2 & si & 0 \leq x < 1 \\ 1 & si & x \geq 1 \end{cases}$$

es la función de distribución que describe la acumulación de probabilidad, tanto de X, como de Y.

Diremos que dos o más variables aleatorias están igual o idénticamente distribuidas si tienen la misma función de distribución.

Siguiendo con el ejemplo anterior, podríamos hallar las funciones de distribución y de probabilidad de la variable X, descrita en el ejemplo 8, calculando $p[X \leq x]$ para todo $x \in R$. Así, por ejemplo, si $0 \leq x < 1$,

$$F_X(x) = p[X \leq x] = p[X \leq 0] = p[[(r,r),(r,n),(n,r),(n,n)]] = \frac{4}{9}$$

Haciendo los cálculos para el resto de valores de x, obtenemos la siguiente función de distribución:

$$F_X(x) = \begin{cases} 0 & si \quad x < 0 \\ 4/9 & si \quad 0 \leq x < 1 \\ 8/9 & si \quad 1 \leq x < 2 \\ 1 & si \quad x \geq 2 \end{cases}$$

cuya gráfica es

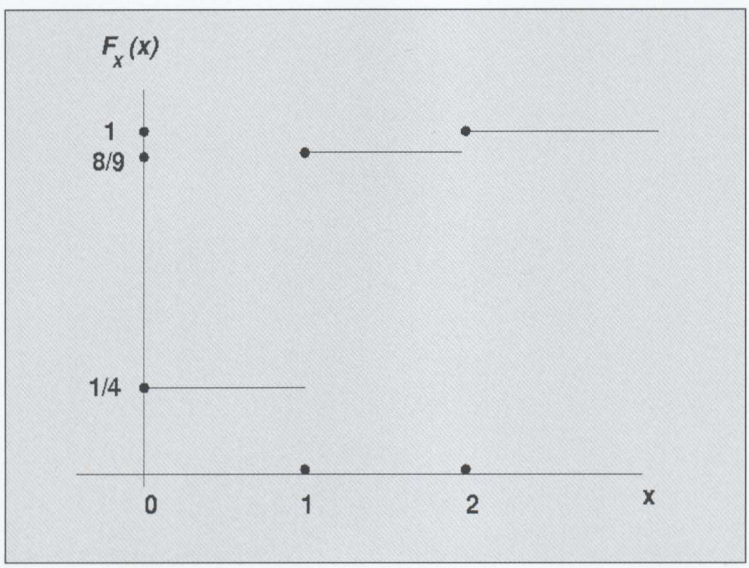

Figura 3.8

Es evidente que esta variable es discreta pues, como ya hemos visto, sus valores posibles son $\{0,1,2\}$. La función de probabilidad para dichos valores es:

$$p[X=0]=p[(r,r),(r,n),(n,r),(n,n)]=\frac{4}{9}$$

$$p[X=1]=p[(b,r),(b,n),(r,b),(n,b)]=\frac{4}{9}$$

y

$$p[X=2]=p[(b,b)]=\frac{1}{9}$$

Resumiendo, la función de probabilidad es

$$p[X=x]=\begin{cases}4/9 & para \quad x=0\\ 4/9 & para \quad x=1\\ 1/9 & para \quad x=2\\ 0 & para \quad resto\end{cases}$$

y su representación gráfica

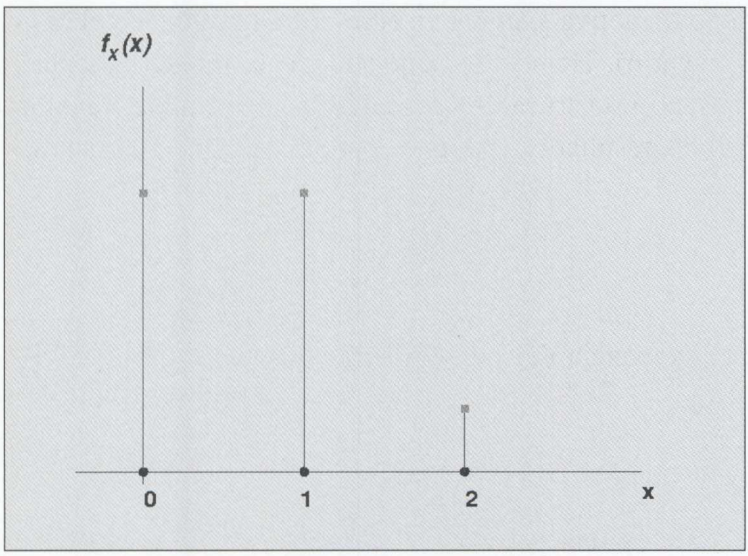

Figura 3.9

Adviértase que el *salto* de la función de distribución en cada uno de los valores de la variable es, precisamente, el valor de la función de probabilidad en dicho punto.

2. ESPERANZA MATEMÁTICA DE UNA VARIABLE ALEATORIA

Es ya conocido que la media aritmética resume a una variable estadística, y si la dispersión (varianza) no es grande, la media es una buena representante de toda la distribución de frecuencias. Cuando vimos esos tópicos en la Estadística Descriptiva, lo justificamos observando que la media era el centro de gravedad de la distribución. Si la media la referimos a una distribución de probabilidad (es decir, cuando tratemos con comportamientos aleatorios) hablaremos de **esperanza matemática, media, valor medio o valor esperado** de una variable aleatoria.

Históricamente, el concepto de esperanza surge de los juegos de azar, al intentar calcular la ganancia que un jugador **esperaba** obtener tras cierto número de partidas.

Así, por ejemplo, lanzamos una moneda y si sale cara, el jugador recibe un euro, perdiéndolo si sale cruz (si te parece que la apuesta no te motiva, cambia un euro por un millón, en el caso de que puedas pagarlo). Hemos de suponer que la moneda es perfecta, es decir, que no está trucada, no siendo más favorable ninguno de los dos posibles resultados. En este caso, la variable se resume como

$$X:\begin{vmatrix} 1 & 1/2 \\ -1 & 1/2 \end{vmatrix}$$

La ganancia del jugador será

$$1\times\frac{1}{2}+(-1)\times\frac{1}{2}=0$$

es decir, espera una ganancia de 0 euros, ganancia nula. Este es el caso de un juego justo o equitativo.

Sea X una variable aleatoria discreta que toma como valores $x_1, x_2, \ldots, x_n$ y con función de probabilidad conocida $p_i = p[X = x_i]$. Se llama esperanza de X al número

$$\mu = E(X) = x_1 \times p_1 + \cdots + x_n \times p_n = \sum_{i=1}^{n} x_i \times p_i$$

Y se llama varianza a

$$\sigma^2 = Var(X) = (x_1 - \mu)^2 \times p_1 + \cdots + (x_n - \mu)^2 \times p_n = \sum_{i=1}^{n} (x_i - \mu)^2 \times p_i$$

siendo su raíz cuadrada positiva, su desviación típica

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

Puede demostrarse que

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} x_i^2 p_i - \mu^2$$

Cuando las variables tomen infinitos valores el sumatorio aparecerá extendido hasta infinito[15].

Recuérdese la similitud existente con la media en variables estadísticas, donde en lugar de p_i escribíamos la frecuencia f_i.

Analicemos otro ejemplo.

Ejemplo 10. En el Instituto, los alumnos del último curso organizan una rifa con 100 papeletas que venden a 2 euros cada una, existiendo un único premio de 100 euros. Un alumno compra 3 boletos, ¿cuál es la ganancia que espera obtener?

Si no obtiene premio, el alumno perderá 6 euros, (3x2), y si le obtiene, ganará 94 euros, (100-3x2), entonces la variable X, *ganancia del jugador*, es

15. La existencia de la esperanza matemática de X viene condicionada porque la serie sea absolutamente convergente, es decir,

$$\sum_{i=1}^{\infty} |x_i| \cdot p_i < \infty$$

Con ello se garantiza la existencia y unicidad de la suma, como serie absolutamente convergente.

$$X:\begin{vmatrix} -6 & p[X=-6] \\ 94 & p[X=94] \end{vmatrix}$$

esas probabilidades son

$$p[X=-6]=p(\overline{P_1}\,\overline{P_2}\,\overline{P_3})=p(\overline{P_1})\times p(\overline{P_2}\mid\overline{P_1})\times p(\overline{P_3}\mid\overline{P_1}\,\overline{P_2})=\frac{99}{100}\times\frac{98}{99}\times\frac{97}{98}=\frac{97}{100}$$

donde P_i designa que *el boleto i-ésimo tiene premio* y $\overline{P_i}$, su complementario. En definitiva, la variable X se expresa como

$$X:\begin{vmatrix} -6 & 97/100 \\ 94 & 3/100 \end{vmatrix}$$

y su ganancia esperada es

$$E(X)=-6\times\frac{97}{100}+94\times\frac{3}{100}=-3$$

es decir, una pérdida media, en este caso, de 3 euros. Este es un ejemplo de juego desfavorable, siendo favorable en el caso de que la esperanza salga positiva.

Su dispersión, es decir, su varianza es

$$Var(X)=E(X^2)-(E(X))^2=\left[(-6)^2\times\frac{97}{100}+(94)^2\times\frac{3}{100}\right]-((-3)^2)=291$$

con una desviación típica de

$$\sigma=17{,}06$$

Es un buen ejercicio intentar descubrir si algún juego de los habituales y comunes en España (1X2, LOTO, Lotería Nacional, ONCE, etc) nos es favorable o desfavorable.

Puedes intentar calcular $E(X)$ y $Var(X)$ en los ejemplos propuestos y desarrollados al comienzo del tema. Por ejemplo, en el

ejemplo 3, donde la variable X es el número de hombres en una familia con tres hijos, su esperanza es

$$\mu = E(X) = 0\times\frac{1}{8}+1\times\frac{3}{8}+2\times\frac{3}{8}+3\times\frac{1}{8}=\frac{12}{8}=1{,}5$$

luego se esperan 1,5 varones en las familias con 3 hijos.

La varianza es

$$\sigma^2 = 0^2\times\frac{1}{8}+1^2\times\frac{3}{8}+2^2\times\frac{3}{8}+3^2\times\frac{1}{8}-(1{,}5)^2=\frac{3}{4}=0{,}75$$

siendo su desviación típica de $\sigma=\sqrt{0{,}75}=0{,}86$.

De todo lo anterior se desprende que la esperanza de una variable aleatoria es un buen resumen de la distribución de probabilidad siempre que no existan valores *raros* de la variable, es decir, muy distantes de la zona donde la variable aleatoria se encuentra mayoritariamente. El siguiente ejemplo clarificará este comentario.

Ejemplo 11. Tras la construcción de una urbanización de gran lujo en un barrio de chabolas, la Concejalía de Asuntos Sociales ha estudiado *el número de hijos por familia* obteniendo la siguiente distribución de probabilidad

$$X:\begin{array}{|cc}0 & 0{,}4\\ 1 & 0{,}3\\ 2 & 0{,}2\\ 23 & 0{,}1\end{array}$$

La esperanza, o número esperado de hijos es

$$E(X)=0\times0{,}4+1\times0{,}3+2\times0{,}2+23\times0{,}1=3$$

luego se esperan 3 hijos en cada familia, pero difícilmente es representativo de una distribución en la que el 90% de las familias tienen, como mucho, 2 hijos.

En este ejemplo, la presencia de un valor de la variable *extremadamente raro* y alejado de la zona donde la variable se encuentra mayoritariamente distorsiona todos los resultados. Planteemos otro ejemplo de complicación mayor, por lo que el alumno puede, y tal vez debe, saltárselo en una primera lectura.

Ejemplo 12. Un fabricante de tornillos afirma que el 5% son defectuosos. Disponemos de una caja de 100 unidades. Estudiemos la variable aleatoria *número de extracciones hasta encontrar el primer tornillo defectuoso*. Parece evidente que los resultados dependerán de que devolvamos, o no, a la caja el tornillo extraído. Planteamos así dos casos, primero, si devolvemos el tornillo extraído, y segundo si no lo hacemos.

Caso 1. Las extracciones se realizan con reemplazamiento, es decir, tras examinar un tornillo, éste se devuelve a la caja.

La variable toma como valores posibles todos los números naturales, con la siguiente función de probabilidad:

$$p[X=1]=p(D_1)=0{,}05$$

$$p[X=2]=p(\overline{D_1}D_2)=p(\overline{D_1})p(D_2)=0{,}95\times 0{,}05$$

.....

$$p[X=k]=p(\overline{D_1}\cdots\overline{D_{k-1}}D_k)=(0{,}95)^{k-1}\times 0{,}05, \qquad k=1,2,\ldots$$

donde D_i indica el suceso *obtener un tornillo defectuoso en la extracción i-ésima* y $\overline{D_i}$, su complementario.

Se puede comprobar si el nivel del curso lo permite, que es una función de probabilidad, es decir, que esas probabilidades están entre 0 y 1 y que su suma vale 1:

$$\sum_{k=1}^{\infty} p[X=k]=\sum_{k=1}^{\infty} 0{,}05\times(0{,}95)^{k-1}=0{,}05\times\frac{1}{1-0{,}95}=1,$$

con representación gráfica

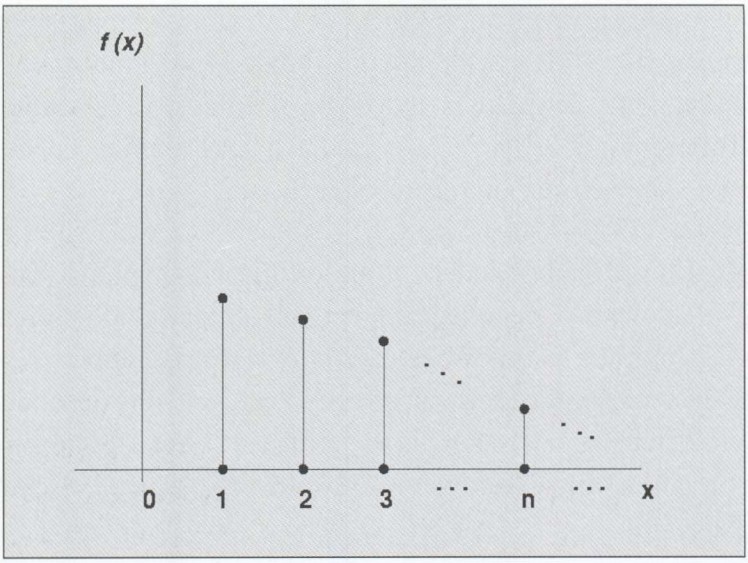

Figura 3.10

La función de distribución es más complicada; facilitamos su expresión

$$F_X(x) = p[X \le x] = \begin{cases} 0 & x < 1 \\ 0{,}05 & 1 \le x < 2 \\ 0{,}05 + 0{,}05 \times 0{,}95 & 2 \le x < 3 \\ \vdots & \vdots \\ 0{,}05 + 0{,}05 \times 0{,}95 + \cdots + 0{,}05 \times 0{,}95^{k-1} & k \le x < k+1 \\ \vdots & \vdots \end{cases}$$

pudiendo expresarse de forma general como

$$F_X(x) = \begin{cases} 1 - 0.95^{k-1} & k \le x < k+1 \quad k = 1,2,\ldots \\ 0 & x < 1 \end{cases}$$

ya que cada valor de la función de distribución es la suma de una progresión geométrica limitada. Su gráfica es:

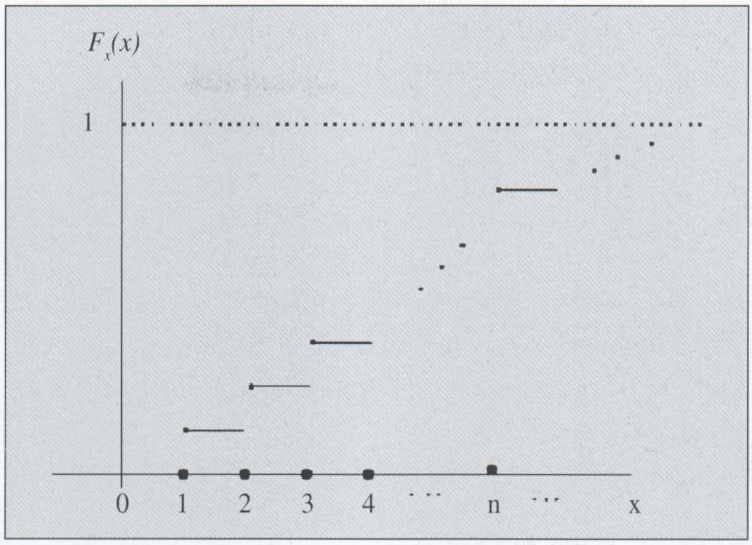

Figura 3.11

A partir de la función de probabilidad, podemos calcular la probabilidad de cualquier suceso. Así, la probabilidad de *realizar al menos 10 extracciones hasta encontrar un tornillo defectuoso* será:

$$p[X \geq 10] = 1 - p[X < 10] = 1 - \sum_{k=1}^{9} p[X = k] = 0{,}95^8 = 0{,}66$$

Asimismo, podemos calcular la probabilidad de un suceso condicionada a la ocurrencia de otro. Por ejemplo, si en las 9 primeras extracciones no ha salido un tornillo defectuoso, podemos calcular la probabilidad de que *no haya que realizar más de 15*. Es decir, una probabilidad condicionada

$$p[X \leq 15 \mid X \geq 10] = \frac{p[X \leq 15, X \geq 10]}{p[X \geq 10]} = \frac{p[10 \leq X \leq 15]}{1 - p[X \leq 9]} = \frac{\sum_{10}^{15} p[X = k]}{1 - \sum_{1}^{9} p[X = k]} = 0{,}26$$

Caso 2. Si las extracciones se realizan sin reemplazamiento.

Ahora, X es una variable aleatoria discreta que toma únicamente los valores 1,2,3,...,96. Su función de probabilidad será:

$$p[X = 1] = p(D_1) = 0{,}05$$

$$p[X=2]=p(\overline{D_1}D_2)=p(\overline{D_1})p(D_2\mid\overline{D_1})=\frac{95}{100}\times\frac{5}{99}$$

.....

$$p[X=k]=\frac{95}{100}\times\frac{94}{95}\times\cdots\times\frac{95-(k-2)}{100-(k-2)}\times\frac{5}{100-(k-1)}$$

tratándose de una función de probabilidad. Lógicamente, y al igual que en el caso anterior, la función de distribución es escalonada, con *saltos* cada vez menores; sin embargo, alcanza el valor uno, en concreto, a partir de $x=96$.

3. MODELOS DE DISTRIBUCIONES

Una vez aproximados a las variables aleatorias, debemos decir que existen muchas variables con "nombre y apellido". Éstas son modelos conocidos con un cierto nombre, que se ajustan adecuadamente a ciertos comportamientos y que podemos utilizar para explicar ciertos fenómenos. Es algo así como si ante la compra de una camisa, vamos a un sastre y nos la hace a medida y con el color que queramos (más tiempo, más dinero, aunque óptima satisfacción) o bien vamos a una tienda y buscamos de entre las de nuestra talla una de un color que nos guste (más rápido, más barato, aunque tal vez el color no sea el deseado o la talla de cuello no sea la exacta)

Ante un cierto fenómeno, podemos intentar descubrir una variable aleatoria que lo explique adecuadamente, o bien, sabidos los fenómenos que explica un cierto modelo de variable, trabajar con él con la consiguiente economía y rapidez.

Abordaremos dos modelos de distribuciones, uno de variable discreta y otro de variable continua; serán las que llamaremos variable aleatoria binomial y variable aleatoria normal.

3.1. Distribución binomial

Frecuentemente, los experimentos consisten en un suceso y su negación. Por ejemplo, al lanzar 5 veces una moneda, o sale cara o su negación, es decir, sale cruz. Elegimos un individuo de una población, o es hombre o es mujer. Elegimos un tornillo, o es defectuoso o no lo es, etc. En cada uno de los ejemplos anteriores y en otros muchos que se te pueden ocurrir, los **sucesos posibles son sólo dos** (c y z, h y m, d y $\overline{d}$, etc.) y **cada nuevo experimento no viene influido por el anterior ni influye en el siguiente**} (la primera tirada de la moneda no influye en la segunda, el sexo del segundo hijo no viene condicionado por el sexo del primero, extraer un tornillo defectuoso no condiciona cómo ha de ser el siguiente que extraigamos, etc.) y, además, la probabilidad del suceso, en cada paso, es siempre la misma ($p(c)=\frac{1}{2}$ en la primera tirada, $p(h)=\frac{1}{2}$ en cada ensayo, $p(d)=0{,}05$ según el ejemplo 12, etc.)

Se denomina *experimento de Bernoulli* a aquél cuya realización da lugar a dos resultados posibles, A o $\overline{A}$, con probabilidades respectivas:

$$p=p(A) \quad y \quad q=1-p=p(\overline{A})$$

Se puede denotar con la notación habitual

$$X:\begin{vmatrix} 1 & si & A & p=p(A) \\ 0 & si & \overline{A} & q=1-p=p(\overline{A}) \end{vmatrix}$$

y se escribe simbólicamente $b(p)$.

Se repite este experimento aleatorio n veces en las mismas condiciones, es decir, con independencia. Cada repetición es, por tanto, independiente de las anteriores y de las posteriores. Sea la variable aleatoria X, *número de veces que sale* A al repetir el experimento n veces. La probabilidad de que ésta tome un valor k es

$$p[X=k]=p(\overbrace{A\ldots A}^{k)}\overbrace{\overline{A}\ldots\overline{A}}^{n-k)})$$

debiendo contar las $\binom{n}{k}$ distintas formas de ordenar los k resultados A y los $n-k$, $\overline{A}$.

En definitiva[16]

$$p[X=k]=\binom{n}{k}p^k(1-p)^{n-k}$$

es la función de probabilidad de una variable aleatoria *binomial* $B(n,p)$, donde n representa el número de veces que se repite de forma independiente el experimento y p la probabilidad de obtener A.

Por ejemplo, si lanzamos una moneda, A es salir cara y $\overline{A}$ salir cruz, $p=\frac{1}{2}$ y $q=1-p=1-\frac{1}{2}=\frac{1}{2}$, lanzar una moneda responde a un modelo Bernoulli $b\left(\frac{1}{2}\right)$. Llamando X al número de caras al lanzar la moneda 1 vez,

$$X:\begin{vmatrix}1 & cara & p=1/2\\ 0 & cruz & q=1-p=1/2\end{vmatrix}$$

Si lanzamos la moneda 5 veces y representamos por X_i al número de caras en el lanzamiento i-ésimo, la suma $X=X_1+X_2+X_3+X_4+X_5$ representará el número de caras al lanzar la moneda 5 veces de forma independiente. Este experimento decimos es una binomial $B(5,1/2)$, el 5 hace referencia al número de lanzamientos y el 1/2 a la probabilidad de cara.

16. La variable aleatoria binomial fue presentada y obtenida por J. Bernoulli en 1713 en su obra *Ars Conjectandi*. Es, en consecuencia, una de las variables aleatorias más antiguas. La variable aleatoria binomial se reconoce habitualmente como el número de éxitos tras la realización de n *tiradas* de Bernoulli de parámetro p, donde la palabra *tiradas* recuerda a los juegos de azar, que son los experimentos aleatorios para los que primero se introdujo esta variable.

La función de distribución de una variable aleatoria binomial es

$$F_X(x) = p[X \le x] = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$$

con lo que en el ejemplo podemos fácilmente calcular la probabilidad de, como mucho, 3 caras al lanzar la moneda 5 veces,

$$p[X \le 3] = p[X=0] + p[X=1] + p[X=2] + p[X=3]$$

pudiendo sustituir y hacer los cálculos pertinentes, o bien, aprender el manejo de una sencillas tablas que proporcionan, como debes comprobar, que esa probabilidad es 0,8125.

En el ejemplo 3, la variable X, número de varones en una familia de 3 hijos, sigue también una distribución binomial. Los sucesos, uno opuesto del otro, en este experimento son:

$$A = \{h\} = hombre \qquad \overline{A} = \{m\} = mujer$$

siendo

$$p = p(h) = \frac{1}{2} \qquad y \qquad q = 1 - p = p(m) = 1 - \frac{1}{2} = \frac{1}{2}$$

Se trata de una $B(3,1/2)$.

La siguiente tabla resume sus probabilidades:

X	$p = p[X = x]$
0	$q^3 = \frac{1}{2^3}$
1	$\binom{3}{1} pq^2 = 3\frac{1}{2}\frac{1}{4} = \frac{3}{8}$
2	$\binom{3}{2} p^2 q = 3\frac{1}{4}\frac{1}{2} = \frac{3}{8}$
3	$p^3 = \frac{1}{2^3}$ 1
	1

Por último, se puede demostrar que la esperanza de una variable aleatoria binomial $B(n,p)$ es

$$E(X)=n\times p$$

y su varianza

$$Var(X)=n\times p\times q$$

lo que nos permite calcular de forma sencilla los valores esperados; así, en el ejemplo 3, una $B(3{,}1/2)$, el número esperado de varones es $n\times p=3\times\frac{1}{2}=1{,}5$, como ya obtuvimos directamente de la aplicación de la fórmula de la esperanza.

Ejemplo 13. El mejor jugador de baloncesto del Instituto tiene aproximadamente un 80% de aciertos de tiros libres. Si en un partido tira 5 tiros libres, ¿cuántos puntos esperamos anotarnos? Trata de describir e identificar la variable. Adicionalmente, puedes tratar de responder a cuestiones como, ¿cuál es la probabilidad de que por lo menos enceste 2?, ¿y de que por lo menos 1?, o ¿que enceste menos de 4?, etc.

3.2. Distribución normal

Finalizamos este tema con un tipo de variable aleatoria continua de uso muy común y generalizado en la práctica estadística[17].

17. Fue en 1733 cuando De Moivre estableció [véase STIGLER, S.M. *The History of Statistics. The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. Cambridge, Mass., 1986. Pág. 70, o una interesante descripción en Daw y Pearson (DAW, R. H. y PEARSON, E.S. *"Studies in the history of probability and statistics. XXX : Abraham de Moivre's 1733 derivation of the normal curve : a bibliographical note."* *Biométrica,* 59. 1972.] la expresión matemática de la distribución normal como una aproximación de la binomial. Posteriormente, Laplace y Gauss la hallaron empíricamente, estudiando la distribución de los errores de medición y convirtiéndose, tras sus trabajos, en la distribución más utilizada en las distintas aplicaciones, de ahí su nombre. En 1774, Laplace (véase LAPLACE, P. S. de. *"Determiner le mimieu que l'on doit prendre entre trois observations données d'un même phénomene."* En *Memoires de Mathématique et Phisique presentées à l'Académie Royale des Sciences par divers Savans,* 6. 1774.) aunque puede consultarse un resumen en Johnson y Kotz (JOHNSON, N. Y KOTZ, S. *Distributions in Statistics: Vol 2: Continuous Univariate Distributions-1.* Wiley. New York, 1970. Pág 45) demostró que es una buena aproximación de la distribución hipergeométrica, obteniendo alrededor del año 1780 unas tablas con valores aproximados de su función de distribución.

La variable toma los valores en toda la recta real, es decir, desde $-\infty$ hasta ∞. Además, esta distribución es simétrica respecto de la media, que coincide con la moda y con la mediana. La figura siguiente aproxima esta idea.

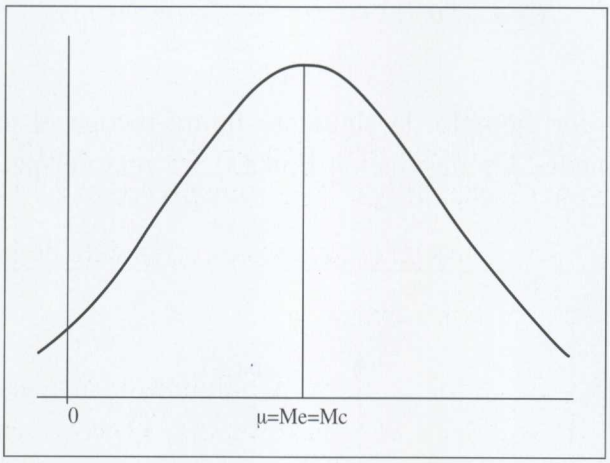

Figura 3.12

Como ya dijimos, en las variables aleatorias continuas no cabe hablar de función de probabilidad sino de lo que hemos llamado **función de densidad**. En concreto, en este caso particular,

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

es la función de densidad de una variable aleatoria normal de media μ y desviación típica σ. Habitualmente se denota como

$$N(\mu,\sigma)$$

con $\mu\in R$ y con $\sigma\in R^+$.

Es evidente que $f(x)>0$, pudiendo comprobarse que la integral en toda la recta real es 1, como exigimos en las condiciones de función de densidad de una variable aleatoria continua.

Tras un proceso denominado de tipificación y consistente en restar a la variable X su media, μ, y dividir el resultado por la des-

viación típica, σ, conseguimos pasar de una $N(\mu,\sigma)$ a una $N(0,1)$ o normal estándar, es decir, una variable cuya distribución aparece centrada en el cero (media o esperanza nula) y varianza la unidad.

$$X \mapsto N(\mu,\sigma) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \mapsto N(0,1)$$

Así, por ejemplo, la siguiente figura recoge el paso de una normal de media 3 y desviación típica 0,5 a una normal estándar.

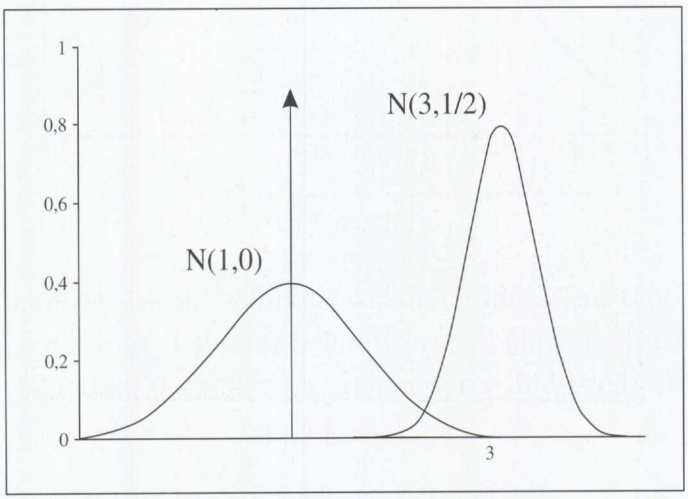

Figura 3.13

La función de densidad correspondiente a este caso es

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Por ejemplo, escribe la función de densidad de una N(3,2) o de una N(1,7) De todos modos, ten presente que esos pares de números no son arbitrarios y que no tendría sentido hablar de una normal acompañada del par (3, -2) ¿Sabes por qué?

La función de distribución en un punto representa la probabilidad *hasta* el punto. En el caso de una variable aleatoria normal, la

función de distribución se representa con Φ. En la gráfica corresponde a la zona sombreada y se calcularía resolviendo la integral siguiente

$$\Phi(z) = p[X \leq z] = \int_{-\infty}^{z} f(t)dt,$$

es decir, calculando el área acumulada bajo la curva y el eje $0X$ desde $-\infty$ hasta z.

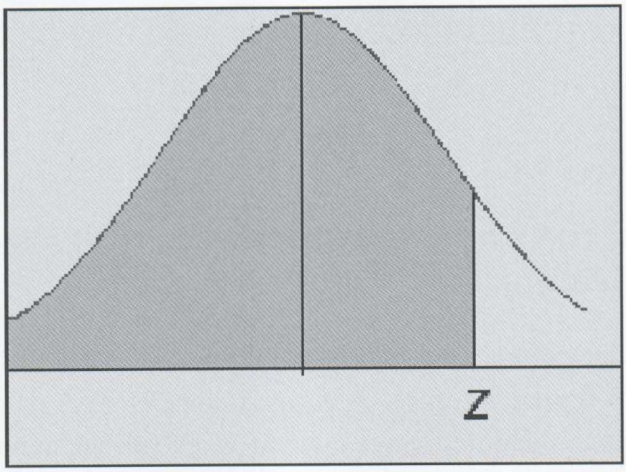

Figura 3.14

Tal vez te hayas percatado de que no puedes resolver esa integral con los métodos habituales. No te preocupes. Existen tablas que proporcionan, dado z, el valor de la función de distribución en él. La única condición es que sólo proporcionan datos para una variable normal estándar, lo que nos obliga a tipificar previamente la variable.

El manejo de estas tablas requiere únicamente darse cuenta de dos cuestiones que se deducen de la definición de la variable aleatoria normal y de las propiedades de la función de distribución.

1. $p[a < X < b] = \int_{a}^{b} f(t)dt = \Phi(b) - \Phi(a)$. Es decir, es el valor de una primitiva en el extremo superior menos en el inferior.

2. Si $z<0$, $\Phi(z)=1-\Phi(-z)$.

Los gráficos de la figura siguiente recogen el sentido de los comentarios anteriores.

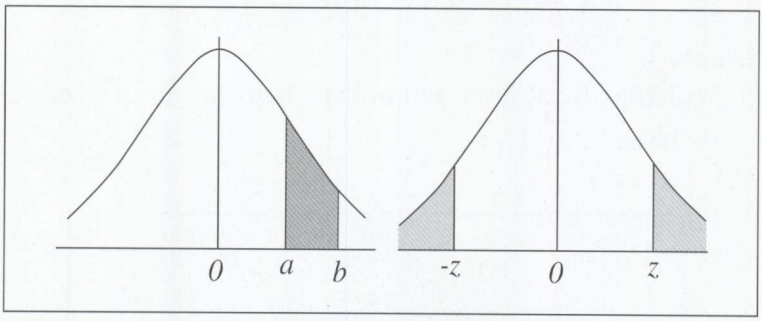

Figura 3.15

4. RELACIÓN ENTRE LAS DISTRIBUCIONES BINOMIAL Y NORMAL

Como dijimos, una variable aleatoria binomial $B(n,p)$ es el resultado de repetir n veces, de forma independiente, un experimento cuya probabilidad de éxito es p. Si n es suficientemente grande (mayor que 20 ó que 30, según autores), nos surge un problema referente a la ausencia de tablas para el cálculo de la distribución de una binomial; por ejemplo, si se lanza una moneda perfecta 30 veces y pretendemos saber cuál es la probabilidad de que el número de caras sea menor que 10, el problema se formaliza diciendo que la variable X, *número de caras*, es una $B(30,1/2)$, pidiéndonos calcular $p[x\leq 10]$. El cálculo es tedioso debido a, como hemos dicho, la ausencia de tablas, recurriendo a la utilización de una aproximación. El histograma de una distribución binomial, como ya hemos visto, puede ser aproximado por una distribución normal. Es necesario que ni p ni q sean próximos a cero, siendo además aceptable la aproximación cuando, o bien

$$p\leq 0{,}5, \quad np\geq 5$$

o bien

$$q\leq 0{,}5, \quad nq\geq 5.$$

Consultando las tablas de la binomial, vemos que no hay datos para $n = 30$. Esto podemos solucionarlo aproximando la distribución $B(n, p)$ a la distribución $N(np, \sqrt{npq})$, es decir, la variable anterior es una $N\left(30 \times \frac{1}{2}, \sqrt{30 \times \frac{1}{2} \times \frac{1}{2}}\right) = N\left(15, \sqrt{7{,}5}\right)$, con lo que la probabilidad pedida es

$$p[X \leq 10] = p\left[\frac{X - 15}{\sqrt{7{,}5}} \leq \frac{10 - 15}{\sqrt{7{,}5}}\right] = \Phi(-1{,}83) = 1 - \Phi(1{,}83) =$$

$$= 1 - 0{,}9664 = 0{,}0336$$

La aproximación se esquematiza de la siguiente forma

$$X \mapsto B(n, p) \xrightarrow{si \quad n>30} X \mapsto N(np, \sqrt{npq})$$

y, recordemos, precisa un tamaño n suficientemente grande y que p no sea demasiado pequeño.

Advirtamos que éste es un problema serio e importante, de hecho, estamos aproximando una variable aleatoria discreta a una variable aleatoria continua.

En el caso de efectuar dicha aproximación se tomará

$$p[X = a] = p[a - 0{,}5 \leq X \leq a + 0{,}5]$$

adviértase que en este caso estamos calculando una probabilidad de una variable aleatoria discreta (la binomial) como la probabilidad ligada a una variable continua (la normal).

Otras aproximaciones útiles son:

$$p[a \leq X \leq b] = p[a - 0{,}5 \leq X \leq b + 0{,}5]$$
$$p[a < X < b] = p[a + 0{,}5 \leq X \leq b - 0{,}5]$$

5. EJERCICIOS PROPUESTOS

Ejercicio 3.1. *Lanzamos una moneda, de la que sabemos que sale cara una vez de cada tres, 100 veces. Calcúlese*

1. *La probabilidad de que el número de caras esté entre 20 y 25.*
2. *La probabilidad de obtener más de 85 caras.*
3. *La probabilidad de que el número de caras esté entre 45 y 55, y la de que sea exactamente 52 caras.*

Responde a las mismas cuestiones si ahora la moneda que se lanza es perfecta y compara los resultados. Comentar si existe alguna evidencia acerca de los resultados.

Ejercicio 3.2. *Últimamente, el profesor de Matemáticas tiene ciertas excentricidades. En su última clase propuso un juego. Se sospecha que el 33% de la población es fumadora; el juego consiste en que el profesor se apostaría en un banco situado a pocos metros del Instituto, observaría a las 10 primeras personas y si al menos la mitad fuman, aprobaría a toda la clase, y si no es así, la suspendería. Tú como alumno, ¿tienes motivos para estar intranquilo?*

No obstante, tras unos breves cálculos, el profesor decide que observará a las 100 primeras personas, aprobando a toda la clase si al menos la mitad fuma, suspendiéndola en caso contrario; ¿tienes ahora motivos para estar más tranquilo?

Ejercicio 3.3. *Compruébese que la primera función es de distribución, no siéndolo las siguientes.*

1. $F_1(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$

2. $F_2(x) = \begin{cases} \frac{2}{3}e^{-x} & x \leq 0 \\ 1 - \frac{2}{3}e^{-x} & x > 0 \end{cases}$

3. $F_3(x)=\begin{cases}0 & x<-1\\ x & |x|\le 1\\ 1 & x>1\end{cases}$

Ejercicio 3.4. *Se lanza dos veces una moneda equilibrada.*

1. *Defínase la aplicación X, número de caras obtenidas.*
2. *Descríbanse los sucesos $[X=1]$ y $[X<2]$, y calcúlense sus probabilidades.*
3. *Hállense las funciones de distribución y de probabilidad de X.*

Ejercicio 3.5. *Un examen tipo test consta de cinco preguntas con tres posibles opciones cada una. Un alumno contesta al azar las cinco cuestiones. Suponiendo que cada respuesta acertada vale dos puntos, hállese la distribución del número de puntos obtenidos por el alumno.*

Ejercicio 3.6. *Resuélvase el ejercicio anterior, en el supuesto de que al alumno se le reste un punto por cada respuesta errónea.*

Ejercicio 3.7. *El número de errores por factura que un contable comete es una variable aleatoria discreta con función de probabilidad:*

$$p[X=x]=e^{-\lambda}\frac{\lambda^x}{x!},\ \text{si}\ \ x=0,1,\ldots,\ (\lambda>0)$$

1. *Calcúlese la probabilidad de que no cometa ningún error.*
2. *¿Cuál es la probabilidad de que cometa algún error?*
3. *Sabiendo que ha cometido al menos un error, ¿cuál es la probabilidad de que no cometa más de cinco?*

Ejercicio 3.8. *Sea X la variable aleatoria que designa el número de coches vendidos cada semana en un establecimiento. Se sabe que X tiene la siguiente función de probabilidad:*

X	0	1	2	3	4	5	6	7	8 ó más
$p[X=x]$	0,04	0,04	K	0,11	0,3	0,23	0,1	0,05	0,03

1. *Hállese el valor de* k.
2. *Determínese la función de distribución de* X.
3. *Calcúlese:* $p[2<X\leq 5], p[X\geq 7]$ y $p[X\leq 6 \mid X>3]$

Ejercicio 3.9. *Un ratón de laboratorio tiene dos posibles caminos para salir de un laberinto. Si toma el primero, caminará 15 centímetros hasta encontrar una bifurcación, siendo 30 y 20 centímetros, respectivamente, las distancias que habrá de recorrer si elige una u otra posibilidad. Si por el contrario se decide por el segundo camino, éste le conducirá a la salida, tras recorrer 60 centímetros. Como el investigador ha impregnado la entrada del segundo camino con un atrayente olor a queso, el ratón elegirá esta opción un 85% de las veces. Hállese la distribución de la variable que expresa la longitud que recorrerá el ratón hasta llegar a la salida.*

Ejercicio 3.10. *Compruébese que* f *no es función de densidad pero que* g *sí lo es.*

1. $f(x)=\begin{cases}3x(x-1)/2 & 0\leq x\leq 2\\ 0 & resto\end{cases}$

2. $g(x)=\begin{cases}1-|1-x| & 0\leq x\leq 2\\ 0 & resto\end{cases}$

Ejercicio 3.11. *Sea* X *una variable aleatoria con función de densidad:*

$$f(x)=\begin{cases}1-|x| & |x|\leq 1\\ 0 & resto\end{cases}$$

1. *Represéntese gráficamente dicha función.*
2. *Compruébese que es función de densidad.*
3. *Calcúlense las siguientes probabilidades:* $p[X\geq 0]$ y $p[|X|<0{,}5]$.

Ejercicio 3.12. *Sea X una variable aleatoria con función de densidad:*

$$f(x)=\begin{cases} 0 & x\le 0 \\ a(1+x) & 0<x\le 1 \\ 2/3 & 1<x\le 2 \\ 0 & x>2 \end{cases}$$

Obténgase:

1. *El valor de a para que f sea función de densidad.*
2. $p[0{,}5 < X \le 1{,}5]$.

Ejercicio 3.13. *Sea X la variable aleatoria que recoge el número de días que un paciente está inscrito en el primer lugar de una lista de espera de la Seguridad Social para ser operado. La función de densidad de X es*

$$f(x)=\begin{cases} \dfrac{1}{30}e^{-\frac{x}{30}} & x>0 \\ 0 & \textit{resto} \end{cases}$$

1. *¿Cuál será el mínimo número de días que tendrá que esperar para ser intervenido con una probabilidad de 0,95?*
2. *Si un paciente lleva inscrito 10 días en el primer lugar de la lista, ¿cuál será la probabilidad de que espere, a lo sumo, tres días más?*

Ejercicio 3.14. *Sea X una variable aleatoria. Compruébese que la siguiente es una función de distribución:*

$$F(x)=\begin{cases} 0 & x<0 \\ \dfrac{2}{3}x & 0\le x<0{,}5 \\ \dfrac{1}{3}(2x+1) & 0{,}5\le x<1 \\ 1 & x\ge 1 \end{cases}$$

BIBLIOGRAFÍA COMENTADA

Fdez-Abascal y otros (FERNÁNDEZ-ABASCAL, H., GUIJARRO, M., ROJO, J. L. Y SANZ, J. A. *Cálculo de Probabilidades y Estadística.* Ariel Economía. Barcelona, 1994) en sus capítulos 4 y 5, detallan formalmente el comportamiento y la distribución de probabilidad de los tipos de variable propuestos, y en sus capítulos 9, 10, 11 y 12, describen, formalizan y estudian todas las variables aleatorias cuyo uso práctico es frecuente.

Asimismo, los manuales Peña (PEÑA, D. *Estadística. Modelos y Métodos 1. Fundamentos.* Alianza Editorial. Madrid, 1991), secciones 3.3 y 3.5; y Fernández y Fuentes (FERNÁNDEZ, C. Y FUENTES, F. *Curso de Estadística descriptiva. Teoría y práctica.* Ariel Economía. Barcelona, 1995), capítulo 6, abordan los contenidos de este capítulo; también lo abordan Casas y Santos (CASAS, J. Y SANTOS, J. *Introducción a la Estadística para economía y administración de empresas.* Centro de Estudios Ramón Areces. Madrid, 1995), capítulos 9, 11 y 13, el clásico Arnáiz (ARNÁIZ, G. *Introducción a la Estadística Teórica.* Lex Nova. Valladolid, 1986), capítulos 2 y 11, o el también clásico Rohatgi (ROHATGI, V. K. *An Introduction to Probability Theory and Mathematical Statistics.* Wiley. New York, 1977) y (ROHATGI, V. K. *Statistical Inference.* Wiley. New York, 1994). Fdez de Trocóniz (FERNÁNDEZ DE TROCÓNIZ, A. *Introducción a las teorías de las Probabilidades. Estadística clásica y Estadística bayesiana.* Autoeditado. Bilbao, 1980), capítulos 5 a 9 y 19, 23, 24 y 26 a 31, plantea en apéndices de los temas, complementos específicos no imprescindibles para continuar con la materia, así como problemas o ejemplos. Wonacott y Wonacott (WONNACOTT, T. H. Y WONNACOTT, R. J. *Introducción a la Estadística.* Limusa. México, 1979), capítulos 4 y 5, de amena y a la vez rigurosa explicación; Hanke y Reitsch (HANKE, J. E. Y REITSCH, A. G. *Estadís-*

tica para negocios. Irwin. Madrid, 1995), capítulo 6; DeGroot (GROOT, M. H. DE. *Probabilidad y Estadística*. Addison-Wesley Iberoamericana. México D.F., 1988), capítulo 3; Walpole y Myers (WALPOLE, R. E. Y MYERS, R. H. *Probabilidad y Estadística*. McGraw Hill. 4ª. ed. México, 1991), capítulo 2, 4 y 5, además de ejemplos y dibujos, acompañan una aplicación de cada variable; Martín-Pliego y Ruiz Maya (MARTÍN PLIEGO, F. J. Y RUIZ MAYA, L. *Estadística I: Probabilidad*. AC. Madrid, 1995), capítulos 2 y 3, y entre el 7 y el 10 describen muchas distribuciones especiales, que también se encuentran bien descritas en Cuadras (CUADRAS, C. M., ECHEVARRÍA, B., MATEO, J. Y SÁNCHEZ, P. *Fundamentos de estadística: Aplicación a las Ciencias humanas*. P.P.U. Barcelona, 1984), capítulo 6. Merece destacar Freund y Simon (FREUND, J. E. Y SIMON, G. A. *Estadística Elemental*. 8a. ed. Prentice-Hall. México, 1994) por sus brillantes comentarios acerca de la esperanza matemática (capítulo 7).

En cuanto a bibliografía referente a la descripción de variables aleatorias concretas, son clásicos, si bien complejos en su estudio, los cuatro volúmenes de Johnson y Kotz (JOHNSON, N. Y KOTZ, S. *Distributions in Statistics: Vol 1: Discrete Distributions*. Wiley. New York, 1969. *Distributions in Statistics: Vol 2: Continuous Univariate Distributions-1*. Wiley. New York, 1970. *Distributions in Statistics: Vol 3: Continuous Univariate Distributions-2*. Wiley. New York, 1971. *Distributions in Statistics: Vol 4: Continuous Multivariate Distributions*. Wiley. New York, 1972), existiendo, asimismo, gran cantidad de manuales por completo dedicados a alguna distribución concreta, por ejemplo, AITCHISON, J. y BROWN, J.A.C. *The log-normal distribution*. Cambridge University Press. Cambridge (UK), 1957.

Aunque son muchos los libros entre los ya citados que contienen gran cantidad de ejercicios resueltos y propuestos, añadamos a esta lista Fdez-Abascal y otros (FERNÁNDEZ-ABASCAL, H., GUIJARRO, M., ROJO, J. L. Y SANZ, J. A. *Ejercicios de cálculo de probabilidades: resueltos y comentados*. Ariel Matemática. Barcelona, 1995), capítulos 3 a 7, específicamente orientado a la resolución de problemas.

Finalmente, sugiramos algunos manuales con enfoque aplicado y que acompañan su tratamiento mediante medios informáticos; así, Mason y Lind (MASON, R. D. Y LIND, D. A. *Estadística para Administración y Economía*. Alfaomega. México, 1992), en sus capítulos 6 y 7, o Levine y otros (LEVINE, D.M., RAMSEY, P.P. y BERENSON, M.L. *Business Statistics for Quality and Productivity*. Prentice-Hall. New Jersey, 1995) secciones 6.4, 6.5 y 6.5. Señalamos especialmente el manual de Freund y Simon (FREUND, J. E. Y SIMON, G. A. *Estadística Elemental*. 8a. ed. Prentice-Hall. México, 1994), capítulos 8 y 9, en el que se plantean ejercicios de simulación; en uno, determinan cómo jugar a cara o cruz sin lanzar realmente una moneda; en el otro, simulan valores de variables aleatorias continuas.

4. ESTIMACIÓN PUNTUAL

1. IDEA GENERAL DE LA INFERENCIA ESTADÍSTICA

La Estadística proporciona técnicas para obtener información acerca de las características de un conjunto más o menos numeroso de individuos que denominamos población [18]. Así, podemos estar interesados en conocer el porcentaje de hogares españoles que disponen de lavavajillas, o el número medio de horas diarias que los alumnos de bachillerato dedican al estudio, o en comparar el efecto de dos medicamentos antitérmicos entre un determinado tipo de pacientes.

Para tener la información que permita conocer las características de una población podemos optar entre las siguientes alternativas:

1. Observar estas características en todos los miembros de la población, disponiendo de esta forma de la información al completo (**censo**).

2. Observar dichas características sólo en una parte significativa de los miembros de la población y, con ciertas precauciones, elevar el resultado obtenido a toda la población (**muestra**).

18. El concepto de población se utiliza en un doble sentido. Por un lado, el conjunto de individuos que se va a analizar: los habitantes de un país, las empresas de una región o los días de un determinado año. Por otro lado, los valores que toma una cierta característica en dichos individuos: el nivel de estudios de los habitantes, el número de empleados de las empresas o la temperatura máxima de cada día. Además, como se ha puesto en evidencia en los tres ejemplos citados, el concepto estadístico de población va más allá de las poblaciones humanas, extendiéndose a cualquier conjunto de personas, animales, cosas,...

Así, para conocer el porcentaje de hogares que cuentan con lavavajillas, podemos encuestar a todos los hogares españoles observando si tienen, o no, este electrodoméstico, o bien, lo que es más habitual, seleccionar adecuadamente un número más o menos reducido de hogares en los que se llevará a cabo la encuesta.

El primer enfoque presenta una ventaja innegable ya que con él se consigue toda la información, lo cual asegurará, en principio, la certeza en todos los resultados obtenidos.

Ahora bien, el enfoque censal presenta también innumerables desventajas, que se tornan consecuentemente en ventajas para la opción muestral:

— La toma de datos en todos los individuos de la población, sobre todo cuando ésta es numerosa, es económicamente gravoso, pues la realización de las observaciones requiere personal humano, medios técnicos, etc. Este inconveniente económico se reduce si la toma de datos se circunscribe a una parte más o menos pequeña de la población.

— Ligado a lo anterior, la opción censal requiere mayor disponibilidad de tiempo que la opción muestral, tanto para la toma de datos como para su posterior procesamiento. A veces, el interés de un dato es inversamente proporcional a la demora en su obtención, y más cuando esta información sirve para tomar una decisión acerca de un cambio de política económica, una estrategia en la comercialización de un producto o un diagnóstico sobre cierta enfermedad.

— El hecho de que la observación censal sea cara y lenta puede producir, si se relajan los requisitos para una toma de datos correcta, una pérdida de calidad en la información, calidad que se consigue con la siempre más cuidadosa observación muestral.

— Ciertas poblaciones, por su propia naturaleza, no admiten ser delimitadas, no pudiendo identificarse todos los indi-

viduos que la forman, requisito necesario para una observación censal. Así, el biólogo que está interesado en el estudio de un cierto tipo de mariposas, nunca podrá disponer de toda la población, teniendo que conformarse con obtener información de una parte de la misma.

— La observación de algunas características de ciertas poblaciones conlleva la destrucción de los individuos de la misma. El especialista en control de calidad de una empresa de neumáticos analizará algunas ruedas sometiéndolas a un rodaje intenso para comprobar su desgaste, lo que inutilizará las mismas para su posterior distribución.

Estas y otras desventajas del método censal y, por tanto, ventajas del método muestral, conducen a inclinarse en la mayoría de las ocasiones por esta última opción. Ahora bien, esta decisión de utilizar una información parcial obliga a tomar una serie de precauciones que van desde una correcta toma de datos (métodos de muestreo)[19], hasta concluir con una interpretación de los resultados en términos de confianza o significación, y no de certeza, pasando por un planteamiento probabilístico de las observaciones.

La *Inferencia estadística*[20] está constituida por un conjunto de técnicas estadísticas que permiten tomar decisiones en ambiente de incertidumbre, es decir, con información parcial.

19. Los llamados métodos de muestreo son un conjunto de procedimientos estadísticos que permiten una adecuada selección de los individuos de la muestra, con el objetivo de que ésta replique, en escala reducida, a la población. Habitualmente, el número de individuos de la población es finito; aún así, por conveniencia (para utilizar modelos matemáticos conocidos), en muchas situaciones supondremos que este número es infinito. En este sentido, hablaremos de muestro en poblaciones finitas y muestreo en poblaciones infinitas, siendo este último el planteamiento bajo el que se desarrolla este capítulo y los dos siguientes. El lector interesado en el muestreo en poblaciones finitas puede consultar CLAIRIN, R. y BRION, P.H. *Manual de muestreo*. 2001.

20. La inferencia estadística, tal y como hoy la entendemos, surge a partir de las distintas aportaciones de Ronald A. Fisher (1890-1960), en especial de su obra *Statistical Methods for Research Workers* publicada en 1925. Estas aportaciones se recogen en FISHER, R.A. *Statistical Methods, Experimental Design and Scientific Inference*. Editado por J.H. Bennett con un prólogo de F. Yates. Oxford University Press. Oxford, 1991.
Fisher, que desde 1919 trabajó en la Estación Experimental de Rothamsted en Inglaterra, desarrolla sus investigaciones ante la necesidad de resolver problemas prácticos relativos

Una vez se ha optado por la opción muestral, ello supone que, para conocer las características de la población en estudio, recabemos información de sólo una parte de los individuos de la misma. De esta forma, encuestaremos a un número razonable de hogares viendo si tienen, o no, lavavajillas, preguntaremos a unos cuantos bachilleres sobre sus horas de estudio, o probaremos el efecto de dos medicamentos antitérmicos en un número reducido de pacientes.

---

a la experimentación agrícola. Como él mismo reconoce en el prólogo de la décimo quinta edición, *"En los años previos a la preparación de este libro, el autor trabajó en estrecha colaboración con los departamentos de investigación biológica en Rothamsted; el libro es decididamente el producto de esta circunstancia. El contacto diario con los problemas estadísticos tal como se planteaban a los trabajadores del laboratorio estimuló las investigaciones puramente matemáticas en las que se basan los nuevos métodos"*.

Sus aportaciones fundamentales a la inferencia, en concreto a la estimación puntual y a los contrastes de hipótesis, se complementaron con el trabajo posterior de Jerzy Neyman (1894-1981) (véase una recensión bibliográfica y científica en el trabajo: *The Neyman-Pearson story: 1926-34. Historical sidelights on an episode in Anglo-Polish collaboration. Festschrift for J Neyman.* New York, 1966, que se presenta en Pearson y Kendall (PEARSON, E.S. y KENDALL, M.G. *Studies in the History of Statistics and Probability* Ed. Griffin. London, 1970. Pág. 455)relativo a los intervalos de confianza. De esta forma, entre los años veinte y treinta del siglo XX se establecieron los tres conceptos básicos y gran parte de los procedimientos operativos de la Inferencia Estadística.

Con todo, las ideas de Inferencia Estadística son anteriores a los trabajos de Fisher y Neyman. Así, por ejemplo, en 1662, J. Graunt (1620-1674) (en MORA CHARLES, Marisol de. *Los inicios de la Teoría de la Probabilidad. Siglos XVI y XVII.* Servicio editorial UPV. Bilbao, 1989. Págs. 185 y siguientes, aparece una buena traducción de la obra básica de Graunt) realiza estimaciones de la población inglesa a partir de una muestra. Sus trabajos abrieron una línea de investigación seguida por W. Petty (1623-1687), que en su obra *Political Arithmetic* publicada en 1690, analiza datos económicos y sociales y por E. Halley (1656-1742), que en 1693, publica una completa tabla de mortalidad (véase Huxley (HUXLEY, G.L., *"The mathematical work of Edmond Halley". Scripta Math*, 24. 1959.) para una descripción de las aportaciones de Halley en este campo o Ayuso y otros (AYUSO, M., CORRALES, H., GUILLÉN, M., PÉREZ-MARÍN, A.M. y ROJO, J.L. *Estadística actuarial Vida.* Ed. Universitat de Barcelona, UB51 manuales. Barcelona, 2001) para una descripción de los contenidos de una tabla de mortalidad desde un enfoque actuarial). En todos estos casos, los fundamentos de la Inferencia no eran estrictamente probabilísticos. Esta fundamentación se incorpora a partir de los trabajos de A. Quetelet (1796-1874), que en 1846, ajusta la estatura de los reclutas a una distribución normal.

Más próximas a la visión actual de la Inferencia Estadística están las aportaciones realizadas en el campo de la Biología y Genética por F. Galton (1822-1911) (remitimos nuevamente a Stigler (STIGLER, S.M. Opus cit. 1986. Pág. 265) y W. Weldon (1860-1906) (véase Pearson y Kendall, (PEARSON, E.S. y KENDALL, M.G. Opus cit. 1970. Pág. 265), encaminadas a contrastar empíricamente la teoría evolucionista de C.R. Darwin (1809-1882). Estos estudios culminan con los trabajos de Karl Pearson (1857-1936) que sistematiza las aportaciones anteriores, promueve nuevos conceptos y procedimientos (entre otros, la prueba ji-cuadrado para contrastar la adecuación entre una distribución observada y una distribución teórica) y aglutina nuevas aportaciones a través de la revista *Biometrika* fundada por él en 1900.

Muchos de los problemas de la Inferencia estadística[21] pueden abordarse suponiendo que la característica a estudiar de la población sigue una distribución conocida. El estudio se centrará, entonces, en el parámetro o parámetros cuyo conocimiento determina completamente la distribución (**Inferencia paramétrica**). Así, podemos suponer que el número de horas de estudio de los bachilleres sigue una distribución normal, planteándonos a partir de ahí, cuál es la media, o si son las chicas más estudiosas que los chicos.

Si no se hace ninguna suposición sobre la distribución de la característica a estudiar, nos limitaremos a enunciar dos hipótesis alternativas acerca de propiedades probabilísticas, aceptando la hipótesis más razonable con la información parcial disponible (**Inferencia no paramétrica**). Por ejemplo, nos podemos plantear si realmente el número de horas de estudio sigue una distribución normal, o si el que un hogar disponga de lavavajillas depende de que la mujer realice o no un trabajo remunerado.

El primer planteamiento, la Inferencia paramétrica, que será el único que desarrollemos en estas unidades, admite tres enfoques complementarios.

— Podemos estar interesados sólo en dar un valor al parámetro desconocido (**Estimación puntual**) Así, trataríamos de dar un valor al número medio de horas de estudio,

21. En este y en los posteriores capítulos desarrollaremos los principales conceptos de la inferencia estadística desde el enfoque clásico. Desde esta perspectiva se supone que la característica poblacional a estudiar (la media, la varianza, la diferencia de medias,...) es fija, no aleatoria; mediante un proceso inductivo a partir de la muestra, trataremos de dar, de un modo u otro, un valor aproximado a dicha característica.
Aunque este enfoque es el más extendido, no es el único. La llamada Inferencia Bayesiana, en honor a Thomas Bayes (1702-1761) (remitimos nuevamente al lector a las referencias Stigler (STIGLER, S.M. Opus cit. 1986. Pág. 88) o Pearson y Kendall (PEARSON, E.S. y KENDALL, M.G. Opus cit. 1970. Pág. 131) que propuso el teorema de la probabilidad inversa, fundamento teórico de este enfoque, establece un esquema diferente. La característica poblacional es aleatoria, disponiéndose de información no muestral sobre la misma (distribución a priori). La información muestral mejorará la información inicial (distribución a posteriori); este proceso podría reiterarse. La distribución final permitirá estimar la característica poblacional en términos probabilísticos. Una visión del enfoque bayesiano se puede ver en DeGroot (DEGROOT, M. H. *Probabilidad y Estadística.* Addison-Wesley Iberoamericana. México, 1988. Capítulos 6 y 7) o en Peña (PEÑA, D. *Fundamentos de Estadística.* Alianza Editorial. Madrid, 2001. Capítulo 9).

siendo este valor coherente con los datos parciales de que dispongamos.

— Nuestro interés puede estar no en dar un valor concreto al parámetro sino en saber entre qué valores se mueve con una cierta confianza en no equivocarnos. (**Estimación por intervalos de confianza**) De este modo, no llegaríamos a decir que el número medio de horas de estudio es, por ejemplo, 2,3, sino que situaríamos ese número entre 2,1 y 2,5, midiendo el posible grado de equivocación.

— Si nuestro propósito es saber si el parámetro toma unos ciertos valores, o no, plantearíamos dos hipótesis que recogieran esta alternativa, optando por la que fuera más razonable de acuerdo con los datos (**Contrastes de hipótesis**) De esta forma, si sabemos que en Francia los bachilleres dedican por término medio 2 horas y media al estudio, podemos tratar de saber si los estudiantes españoles son más o menos aplicados que los franceses.

En esta unidad desarrollaremos la Estimación puntual, dejando los otros dos enfoques para posteriores capítulos.

2. ESTIMACIÓN PUNTUAL: PLANTEAMIENTO

Un ejemplo nos servirá para introducir este primer tipo de Inferencia.

Ejemplo 1. Queremos conocer el porcentaje de hogares españoles con lavavajillas. Descartada la opción censal, en base a las consideraciones anteriores, se elige una muestra representativa del conjunto de hogares (sin entrar ahora en los métodos de muestreo, supondremos que la elección de esa muestra es correcta). En concreto, elijamos una muestra de tamaño 1000. De los 1000 hogares encuestados, 230 poseían lavavajillas, mientras que 770 carecían de él; esto es, en el 23% de los hogares encuestados había este electrodoméstico. No parece arriesgado, si no se posee otra información, asignar este por-

centaje a toda la población. Así, sin tener especiales conocimientos estadísticos, se puede decir que el 23%, más o menos, de los hogares españoles tiene lavavajillas.

Trataremos de formalizar, en lo posible, el problema para dar a la anterior afirmación una cierta consistencia. De entrada, la presencia de lavavajillas en un hogar se resume con una variable de Bernoulli:

$$X = \begin{cases} 1 & \text{Hogar con lavavajillas} \\ 0 & \text{Hogar sin lavavajillas} \end{cases}$$

Además, la probabilidad de que un hogar elegido al azar disponga de lavavajillas, esto es, $p[X=1]$, será p, la proporción de hogares de toda la población que cuenta con dicho electrodoméstico:

$$\text{p[Hogar con lavavajillas]} = p[X=1] = p$$
$$\text{p[Hogar sin lavavajillas]} = p[X=0] = q = 1\text{-}p$$

El valor de p, entre 0 y 1, es desconocido. Si se observara toda la población, p podría determinarse con certeza, mas como tenemos información parcial, nos conformaremos con un conocimiento aproximado y sujeto a incertidumbre. En primer lugar, determinaremos un valor *próximo* al verdadero valor de p, valor que constituirá una estimación del parámetro, $\hat{p}$.

La *Estimación puntual*[22] es el primer escalón de la Inferencia paramétrica que permite dar al parámetro de la distribución un valor que se aproxime al verdadero valor desconocido del mismo.

Para estimar la proporción de hogares con lavavajillas, tomaremos la información de un número más o menos grande de hogares, n, esto es, haremos n observaciones, siendo n, tamaño de la muestra,

22. Como ya se ha comentado anteriormente, los conceptos básicos de Estimación puntual se establecieron en el libro de R.A. Fisher (1925) (véase nuevamente FISHER, R.A. *Statistical Methods, Experimental Design and Scientific Inference*. Editado por J.H. Bennett con un prólogo de F. Yates. Oxford University Press. Oxford, 1991) que sistematiza y da forma a aportaciones anteriores y propone nuevos conceptos (eficiencia, suficiencia,...).

menor que N, tamaño de la población[23]. Cada observación se formaliza mediante una variable de Bernoulli:

$$X_i = \begin{cases} 1 & \text{Si el hogar i-ésimo dispone de lavavajillas} \\ 0 & \text{Si el hogar i-ésimo no dispone de lavavajillas} \end{cases}$$

Las variables $(X_1,\ldots, X_n)$ reciben el nombre de **muestra aleatoria simple**. La observación realizada, que proporcionó 230 hogares con lavavajillas y 770 sin él, se formaliza como una sucesión de 230 unos y 770 ceros en un cierto orden, en principio irrelevante. Por ejemplo, (1,0,0,1,…,1), es decir, el primer hogar tiene lavavajillas, el segundo y tercero no, el cuarto sí,..., el último hogar tiene lavavajillas.

El problema de estimar una proporción puede entenderse mejor mediante un esquema de urnas y bolas. Consideramos toda la población como una urna con N bolas, N_1 blancas (hogares con lavavajillas) y N_2 negras (hogares sin él)

La proporción de bolas blancas, $p = \frac{N_1}{N}$, es desconocida y constituye el objeto de nuestro estudio. Para obtener información acerca de p, en principio y por sencillez expositiva, extraemos con reemplazamiento n bolas. De estas n bolas resultan n_1 blancas y n_2 negras, siendo $\hat{p} = \frac{n_1}{n}$ la proporción de bolas blancas en la muestra.

Este planteamiento de urna admite ser reproducido en clase con los alumnos; se pueden hacer sucesivas extracciones y comprobar empíricamente que una misma estructura de urna produce distintas muestras, incidiendo de esta forma en la aleatoriedad de las mismas.

La suposición de que la muestra puede ser una buena representación de la población es la idea subyacente en toda la Estadística inferencial. En este sentido, las características muestrales pueden dar

23. La consideración de la muestra como aleatoria, esto es, antes de ser observada, es la clave de todo el desarrollo de la Inferencia estadística.

una idea aproximada de las correspondientes características poblacionales.

Entonces, como nuestro interés es conocer la proporción de hogares con lavavajillas, p, nos conformaremos con conocer esta proporción entre los hogares de la muestra, esto es, la proporción muestral, $\hat{p}$. Esta proporción muestral, número de hogares con lavavajillas sobre el número total, no es otra cosa que la media muestral de $(X_1, X_2, \ldots, X_n)$:

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n} = \overline{X}.$$

Para nuestra observación, esta proporción toma el valor 0,23.

El siguiente cuadro recoge los aspectos relativos tanto a la formalización de nuestro problema como a su *visualización* mediante un esquema de urna:

$X = \begin{cases} 1 & \text{Hogar con lavavajillas} \\ 0 & \text{Hogar sin lavavajillas} \end{cases}$	$p[X=1]=p$ $p[X=0]=1-p$
Objeto de estudio	p (proporción poblacional)
Muestra aleatoria simple	$(X_1, X_2, \ldots, X_n)$
Realización de la muestra	$(1, 0, 0, 1, \ldots, 1)$
Proporción muestral	$\hat{p} = \frac{X_1 + \cdots + X_n}{n} = \overline{X}$
Realización de la prop. muestral	$\hat{p} = \frac{1+0+0+1\cdots+1}{1000} = 0{,}23$

3. LA DISTRIBUCIÓN DE LA MUESTRA

Pero la anterior estimación, ¿es buena? Para tratar de dar alguna respuesta es imprescindible considerar la muestra antes de ser observada, es decir, tratarla como variable aleatoria de dimensión n. Para que se puedan seguir los desarrollos con facilidad, echando luz sobre los conceptos que hay detrás de los mismos, supongamos que

la muestra elegida es de tamaño 3, totalmente insuficiente para tomar decisiones para una población grande, pero muy conveniente para nuestro propósito expositivo.

Así, consideraremos que vamos a realizar las observaciones en tres hogares, es decir, tres extracciones de la urna:

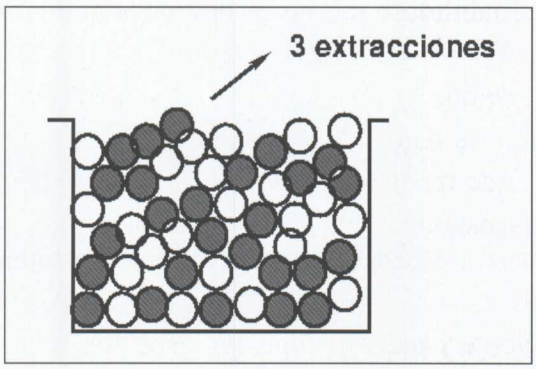

Figura 4.1

Al hacer tres observaciones nos podemos encontrar con muy distintos resultados: los tres hogares sin lavavajillas, sólo uno con lavavajillas, etc. En este sentido, la muestra, como variable aleatoria, presenta unos posibles valores con una probabilidad que podemos calcular:

$(X_1, X_2, X_3) =$		
	(0, 0, 0)	$(1-p)^3$
	(1, 0, 0)	$p\ (1-p)^2$
	(0, 1, 0)	$p\ (1-p)^2$
	(0, 0, 1)	$p\ (1-p)^2$
	(1, 1, 0)	$p^2\ (1-p)$
	(1, 0, 1)	$p^2\ (1-p)$
	(0, 1, 1)	$p^2\ (1-p)$
	(1, 1, 1)	p^3

Trataremos de explicar el contenido de la tabla anterior. Así, por ejemplo, la terna (1, 0, 1) indica que el primer y tercer hogar encuestado tienen lavavajillas y el segundo no, o lo que es lo mismo, volviendo al esquema de urna, que se ha extraído una bola blanca en la primera y tercera extracción, y una negra en la segunda. Como las

extracciones se realizan con reemplazamiento y en cada nueva extracción la urna presenta la misma estructura, la probabilidad de esta terna resulta:

$$p\,[(X_1,X_2,X_3)=(1,\,0,\,1)] = p\,[B_1,N_2,B_3] = p\,[B_1]\,p[N_2]\,p[B_3] = p^2\,(1-p).$$

De la misma forma se explican el resto de valores de la variable y sus probabilidades.

Obsérvese que la consideración de la muestra como un determinado conjunto de datos se ha transformado en muestra como variable aleatoria que recoge todas las muestras que pudiéramos obtener, con sus correspondientes probabilidades.

4. RESUMIENDO LA INFORMACIÓN: ESTADÍSTICOS Y ESTIMADORES

Toda la información sobre p está en la muestra. Así, si se obtiene la muestra (0, 0, 0), indicaría que la mayor parte de los hogares no dispone de lavavajillas, o, en términos de urna, que ésta contiene casi todas las bolas negras. Si se obtiene la muestra (1, 0, 0), (0, 1, 0) o (0, 0, 1), indicaría que una proporción relativamente pequeña de hogares dispone de lavavajillas (la urna está compuesta por una proporción relativamente grande de bolas negras) De la misma forma, explicaríamos la información sobre p contenida en las muestras (1, 1, 0), (1, 0, 1) y (0, 1, 1), o en la muestra (1, 1, 1)

Aunque la muestra contiene toda la información sobre p, su dimensión la hace poco manejable, además de que no permite concretar un valor de p como es nuestro propósito. Para salvar este inconveniente, introducimos los **estadísticos**, funciones de la muestra que, a su vez, son variables aleatorias, y una clase especial de los mismos, los **estimadores**, estadísticos que toman sus valores entre el conjunto de posibles valores del parámetro; esta característica permitirá, una vez concretada la muestra, dar un valor estimado del parámetro, que será el valor concreto del estimador. Sugerimos tres esta-

dísticos, la suma $S = X_1+X_2+X_3$, el producto, $P = X_1 X_2 X_3$, y la media muestral, $\overline{X} = \frac{X_1 + X_2 + X_3}{3}$, cuyas distribuciones vienen recogidas en la tabla siguiente:

S	*Probabilidad*	P	*Probabilidad*	$\overline{X}$	*Probabilidad*
0	$(1-p)^3$	*0*	$1-p^3$	*0*	$(1-p)^3$
1	$3p\ (1-p)^2$	*1*	p^3	*1/3*	$3p\ (1-p)^2$
2	$3p^2\ (1-p)$			*2/3*	$3p^2\ (1-p)$
3	p^3			*1*	p^3

Sólo el estadístico producto y el estadístico media muestral son estimadores, pues toman valores entre 0 y 1, posibles valores de *p*. Ahora bien, mientras que la media muestral conserva toda la información que sobre *p* contenía la muestra, el estimador producto, en la reducción de la dimensión que supone el paso de la muestra al estadístico, ha perdido parte de dicha información. Así, si la media muestral es 2/3, induciremos que en la población hay muchos hogares con lavavajillas (la urna estaría compuesta por un número relativamente grande de bolas blancas frente a un número relativamente pequeño de bolas negras). Esta información es la misma que nos ofrecen los resultados muestrales (1, 1, 0), (1, 0, 1) y (0, 1, 1). Si sólo conocemos el valor de la media muestral y desconocemos de qué realización de la muestra provenía, la única pérdida de información se refiere al orden de salida de las dos bolas blancas y de la bola negra, orden irrelevante de cara a conocer el verdadero valor de *p*.

En cambio, si el estimador *P* vale cero, esto nos puede hacer pensar que se han observado muestras tan dispares como la (0, 0, 0) o la (1, 1, 0) entre otras. En el primer caso, induciríamos una población constituída mayoritariamente por hogares sin lavavajillas (una urna con casi todas las bolas negras), mientras que en el segundo caso, induciríamos una población con una proporción relativamente alta de hogares con dicho electrodoméstico (una urna con bastantes bolas blancas y pocas negras). Por tanto, el resumen de la muestra que proporciona el estimador *P* conlleva una pérdida de información.

La comparación anterior se debería extender a todos los estimadores de p, de hecho infinitos. Admita el lector que esta comparación sería siempre favorable a la media muestral, estimador que conserva toda la información que sobre p contenía la muestra (**estadístico suficiente**). Entonces, parecería sensato elegir dicho estimador para tratar de acercarnos al verdadero valor de p.

Esta elección se ratifica por el hecho, ya comentado, de que tratamos de estimar una proporción poblacional y lo razonable es hacerlo con la proporción muestral.

$$\overline{X} = \frac{X_1 + X_2 + X_3}{3} = \frac{\text{número de bolas blancas}}{\text{número de bolas extraídas}} = \text{proporción de bolas blancas en la muestra.}$$

Elegido el estimador de p, $\hat{p} = \overline{X}$, concretaremos la muestra, lo que nos llevará a concretar la estimación. Realizada la observación, obtuvimos que el primer hogar tenía lavavajillas, el segundo y el tercero no; esto es, en el esquema de urna, se extrajeron una bola blanca, una negra y una negra. Entonces:

Muestra aleatoria simple	*Realización de la muestra*
(X_1, X_2, X_3)	$(1, 0, 1)$
Estimador	*Estimación*
$\hat{p} = \overline{X}$	$\hat{p} = \frac{1+0+1}{3} = 0{,}66$

Por tanto, de acuerdo con la información obtenida, estimamos que el 33,3% de los hogares españoles dispone de lavavajillas.

5. ALGUNAS PROPIEDADES DESEABLES DE LOS ESTIMADORES

Insistamos en algo que, en principio, puede ser sorprendente: un estimador, en este caso la media muestral, no es en principio un valor concreto, sino que es una variable con su distribución, en este caso:

$$\begin{array}{c|l} \overline{X} = & \\ 0 & p\,[\overline{X}=0] = (1-p)^3 \\ 1/3 & p\,[\overline{X}=1/3] = 3p\,(1-p)^2 \\ 2/3 & p\,[\overline{X}=2/3] = 3p^2\,(1-p) \\ 1 & p\,[\overline{X}=1] = p^3 \end{array}$$

Como variable que es, podemos calcular alguna de sus principales características: la esperanza o media teórica y la varianza. Su cálculo es sencillo pues, al tratarse, en este caso, de variables discretas, estas características siguen un desarrollo similar a las características empíricas de la estadística descriptiva, sin más que cambiar frecuencias por probabilidades.

Por tanto, la media o esperanza de $\overline{X}$ vale:

$$E(\overline{X}) = 0\cdot(1-p)^3 + \frac{1}{3}\cdot p(1-p)^2 + \frac{2}{3}\cdot 3p^2(1-p) + 1\cdot p^3 = p.$$

La media de la media muestral (puede asombrar su propio enunciado) es p, el valor que queremos conocer. Téngase en cuenta que cuando utilizamos un estimador, nos equivocaremos en mayor o menor medida, pero utilizando el estimador media muestral al menos aseguramos que no nos equivocamos por término medio o, en otras palabras, el valor esperado es el buscado. Un estimador cuya esperanza coincida con el parámetro a estimar se dice que es un **estimador insesgado**.

De la misma forma, la varianza de $\overline{X}$ será:

$$Var(\overline{X}) = (0-p)^2\cdot(1-p)^3 + (\frac{1}{3}-p)^2\cdot p(1-p)^2 + (\frac{2}{3}-p)^2\cdot 3p^2(1-p) + (1-p)^2\cdot p^3 =$$

$$= \frac{p(1-p)}{3}.$$

Si comparásemos esta varianza con la de cualquier otro estimador insesgado, comprobaríamos que $\overline{X}$ es el de menor varianza, denominando a un estimador que cumpla esta propiedad **estimador insesgado de mínima varianza** o, bajo ciertas condiciones, **estimador eficiente**. Además, el estimador $\overline{X}$ presenta otra buena propiedad ligada a su distribución. En la medida en que aumentamos el

tamaño muestral, los valores del estimador se irán concentrando alrededor del verdadero valor del parámetro, entendiéndose que si la muestra pudiera ser infinita, el estimador sería una constante, siendo esta constante el verdadero valor de p (**estimador consistente**)

Por tanto, siempre que nos planteemos estimar la proporción en la que se presenta una característica entre los miembros de la población, recurriremos a la media muestral $\overline{X}$, pues cumple cuatro propiedades deseables en cualquier estimador: suficiencia, insesgadez, mínima varianza y consistencia.

Además, este estimador es el más razonable en el sentido del **criterio de la máxima verosimilitud**, criterio fundamental en cualquier planteamiento inferencial y método rápido y sencillo para encontrar directamente buenos estimadores.

6. OBTENCIÓN DE ESTIMADORES: EL CRITERIO DE LA MÁXIMA VEROSIMILITUD

El planteamiento anterior de buscar por inspección un buen estimador no parece ser operativo. Para evitar esto, la Estadística cuenta con ciertos procedimientos que, diseñados con criterios razonables, permiten encontrar directamente buenos estimadores. El criterio más utilizado es el de la máxima verosimilitud[24].

El criterio de la máxima verosimilitud se fundamenta en la siguiente idea: Ha ocurrido un cierto suceso, por ejemplo, en nuestro caso, (1, 0, 0). La probabilidad de este suceso depende del parámetro a estimar,

$$\mathbf{p}\ [(X_1, X_2, X_3) = (1, 0, 0)] = p(1-p)^2,$$

24. En el mismo contexto de estimación (estimar el parámetro de una distribución a partir de muestras aleatorias simples), otro método de interés es el de los momentos. Según él, se estiman las características poblacionales (media poblacional, varianza poblacional, mediana poblacional, coeficiente de correlación poblacional,...) con las correspondientes características muestrales; además, si el parámetro a estimar fuese una función de características poblacionales, su estimador sería la misma función de las correspondientes características muestrales.
En otros contextos de estimación cabe señalar los métodos de los mínimos cuadrados y de la c^2 mínima. Para mayor detalle de estos tres métodos se puede consultar Ruiz-Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. J. *Estadística II: Inferencia*, AC. Madrid, 1995. Cápitulo 7).

probabilidad que tomará distintos valores según sea p. Si ha ocurrido (1, 0, 1), supondremos que tenía gran probabilidad de ocurrir, la máxima posible. Buscaremos el valor de p que haga máxima esa probabilidad. El valor hallado será la estimación máximo verosímil de p para ese suceso.

El problema se concreta en maximizar la función de p que proporciona la probabilidad del suceso, esto es, la **función de verosimilitud**:

$$\max_p p[(X_1, X_2, X_3) = (1,0,1)] = \max_p p^2(1-p).$$

Este problema de optimización se puede abordar gráficamente (representando sobre un sistema de ejes cartesianos la función de verosimilitud), por aproximaciones sucesivas (hallando la función de verosimilitud para distintas mallas de valores de p que vayan acercándose al valor óptimo), o derivando la función de verosimilitud e igualándola a cero. En este último caso, resulta:

$$\frac{d}{dp} p[(X_1, X_2, X_3) = (1,0,1)] = \frac{d}{dp}\left[p^2(1-p)\right] = 2p - 3p^2 = 0,$$

siendo $p = 1/3$ un valor que anula esta derivada. Hallando la segunda derivada en dicho punto,

$$\frac{d^2}{dp} p[(X_1, X_2, X_3) = (1,0,1)]\Big|_{p=2/3} = 2 - 6p\Big|_{p=2/3} = -2 < 0,$$

se comprueba la condición de máximo. Por tanto, $\hat{p} = 2/3$ es la estimación máximo verosímil de p cuando se ha observado (1, 0, 0), donde este valor no es otra cosa que la proporción muestral.

El cuadro siguiente resume el planteamiento y resultado de la estimación de una proporción.

PROBLEMA: **Estimar la proporción p con que se presenta la característica A entre los individuos de una población**

$$X = \begin{cases} 1 & \text{ocurre A} \quad p[X=1] = p(A) = p \\ 0 & \text{no ocurre A} \quad p[X=0] = p(\overline{A}) = 1-p \end{cases}$$

$X \to B(p)$ $\quad p$ desconocido, $\quad 0 \le p \le 1$

Se realizan n observaciones: $(X_1, \ldots, X_n)$

Estimador de p: $\boxed{\hat{p} = \overline{X}}$

La proporción muestral es: suficiente, insesgada, de mínima varianza, consistente y máximo verosímil

7. OTROS PROBLEMAS DE ESTIMACIÓN PUNTUAL

El segundo problema que nos plantearemos es hacer inferencia sobre los parámetros de una variable cuantitativa, más en concreto, sobre su media y su varianza.

Ejemplo 2. El Ministerio de Educación y Ciencia está interesado en conocer las horas diarias que dedican al estudio los bachilleres españoles. Para ello, trata de estimar el número medio de horas diarias que dedican al estudio, contando únicamente con la información que proporciona una muestra representativa.

En la mayor parte de estas situaciones supondremos que la variable en cuestión sigue una distribución normal de parámetros μ (media poblacional) y σ (varianza poblacional) desconocidos, aunque para la mayoría de los resultados que comentaremos a continuación, esta suposición de normalidad es irrelevante:

X: número de horas diarias de estudio de un alumno de bachillerato

$X \to N(\mu, \sigma)$ siendo μ y σ desconocidos.

Para decidir acerca de μ, tomemos de nuevo una muestra aleatoria simple $(X_1, \ldots, X_n)$. Parece razonable que si queremos saber acerca de la media poblacional μ, utilicemos la media muestral

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

pudiéndose probar, mediante desarrollos más complejos que los utilizados al estudiar la media muestral como estimador de una proporción, que es realmente un buen estimador, en el sentido de que es suficiente, insesgado, de mínima varianza y consistente, siendo además, el estimador que se obtendría aplicando el método de la máxima verosimilitud. Así, dado que las depauperadas arcas del Ministerio sólo dieron para realizar la encuesta a 10 estudiantes, los resultados obtenidos en cuanto al número de horas diarias de estudio de cada uno de ellos fueron:

1.5 2 0.8 1.3 2.5 1.1 0.3 0 0.6 2.4

Si calculamos la media de estos valores,

$$\overline{X} = \frac{1.5 + 2 + \cdots + 2.4}{10} = 1.25$$

podemos inducir, con mucho riesgo dada la cortedad de la muestra, que los bachilleres españoles dedican una media diaria de 1.25 horas al estudio.

El Ministerio también está interesado en conocer la varianza de la variable *horas diarias de estudio*, entre otros motivos, por saber la representatividad del resultado anterior. Al abordar la estimación de la varianza σ^2, debemos recurrir, por lógica, a la varianza muestral,

$$S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}.$$

Ahora bien, este estimador no es insesgado para estimar σ^2, en concreto,

$$E(S_X^2) = \frac{n-1}{n}\sigma^2,$$

si bien para un n grande, prácticamente desaparece el sesgo (estimador asintóticamente insesgado). Este inconveniente, de cierto relieve cuando trabajamos con muestras pequeñas, lleva introducir un estimador alternativo, la *cuasivarianza muestral*:

$$S_c^2 = \frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1},$$

estimador que, cuando n es grande, tomará valores muy próximos a los de la varianza muestral ya que

$$S_c^2 = \frac{n}{n-1}S_X^2.$$

Ahora, este nuevo estimador sí es insesgado,

$$E(S_c^2) = \sigma^2,$$

siendo conveniente su utilización cuando el tamaño muestral es pequeño.

Por lo demás, ambos son suficientes y consistentes, siendo la cuasivarianza el estimador insesgado de mínima varianza. El método de máxima verosimilitud proporciona como estimador la varianza muestral.

En nuestro caso, dado que el tamaño muestral es pequeño, tenemos que recurrir a la cuasivarianza como estimador de la varianza poblacional:

$$\hat{\sigma}^2 = S_c^2 = \frac{(1.5-1.25)^2 + \cdots + (2.4-1.25)^2}{9} = 0.7361.$$

El cuadro siguiente recoge un resumen del planteamiento de la estimación muestral para poblaciones normales.

PROBLEMA: **Estimar la media y la varianza de una variable con distribución normal**

$X \to N(\mu, \sigma)$	μ y σ desconocidos, $-\infty \leq \mu \leq \infty$, $0 < \sigma < \infty$
Se realizan n observaciones:	$(X_1, \ldots, X_n)$
Estimador de μ:	$\boxed{\hat{\mu} = \overline{X}}$
La media muestral es:	suficiente, insesgada, de mínima varianza, consistente y máximo verosímil
Estimador de σ^2:	$\hat{\sigma}^2 = \begin{cases} S_X^2 & \text{varianza muestral si n grande} \\ S_c^2 & \text{varianza muestral si n pequeño} \end{cases}$
$\boxed{\hat{\sigma}^2 = S_X^2}$:	suficiente, sesgado, consistente y máximo verosímil
$\boxed{\hat{\sigma}^2 = S_c^2}$:	suficiente, insesgado, de mínima varianza y consistente.

Al hilo del primer ejemplo, podríamos estar interesados en comparar el porcentaje de hogares con lavavajillas en dos regiones, problema que se visualiza con dos urnas con bolas blancas y negras en distinta composición. Así, si p_1 es la proporción de hogares con lavavajillas en la primera región y p_2 es la proporción de hogares con lavavajillas en la segunda, la diferencia entre p_1 y p_2, p_1-p_2, nos servirá para evaluar si la primera región presenta una mayor, menor o igual dotación de este electrodoméstico, según que p_1-p_2 sea positivo, negativo o nulo. Ahora bien, de nuevo tanto p_1 como p_2 son desconocidos, teniendo que recurrir a sus respectivos estimadores. Por ello, tomando una muestra de tamaño n_1 de la primera población y de tamaño n_2 de la segunda, las correspondientes proporciones muestrales serán unos buenos estimadores de p_1 y p_2,

$$\boxed{\hat{p}_1 = \overline{X}_1 \qquad \hat{p}_2 = \overline{X}_2},$$

y, entonces, utilizaremos como estimador de p_1-p_2, la diferencia de proporciones muestrales:

$$\widehat{p_1 - p_2} = \bar{X}_1 - \bar{X}_2$$

Igualmente, si queremos comparar las horas diarias de estudio de los bachilleres de colegios públicos y de colegios privados, con sólo información parcial, y llamando μ_1 al número medio de horas de estudio en los colegios públicos y μ_2 al número medio en los colegios privados, recurriremos al estimador diferencia de medias muestrales:

$$\widehat{\mu_1 - \mu_2} = \bar{X}_1 - \bar{X}_2 .$$

Si el interés fuera comparar la dispersión de las horas de estudio en los dos tipos de enseñanza, esto es, comparar σ_1^2 y σ_2^2 nos plantearemos la estimación de su cociente, σ_1^2 / σ_2^2. Siguiendo la misma lógica, utilizaremos como estimador de este cociente

$$\frac{S_{X_1}^2}{S_{X_2}^2} \quad \text{o} \quad \frac{S_c^2}{S_{c_2}^2},$$

según contemos con muestras suficientemente grandes o no.

Ejemplo 3. La Concejalía de Cultura y Educación del Ayuntamiento ha efectuado una encuesta para conocer la situación cultural de las familias de la ciudad. Para ello ha encuestado a 20 familias elegidas al azar tomando nota, entre otras cosas, del número de libros de literatura existentes en el hogar y del nivel de estudios del padre (**E**studios **P**rimarios, **E**studios **M**edios, **E**studios **S**uperiores). Dichos datos se recogen en la tabla siguiente:

Familia	*Número de libros*	*Nivel de estudios*
1	87	E.S.
2	52	E.P.
3	150	E.S.
4	62	E.M.
5	93	E.M.
6	15	E.P.
7	43	E.M.

Familia	*Número de libros*	*Nivel de estudios*
8	52	E.M.
9	23	E.P.
10	79	E.M.
11	104	E.S.
12	90	E.M.
13	53	E.P.
14	31	E.M.
15	12	E.P.
16	101	E.M.
17	240	E.S.
18	153	E.S.
19	79	E.M.
20	5	E.P.

1. Estimar el número medio de libros en los hogares de la ciudad.
2. Estimar la proporción de familias en las que el padre tiene estudios primarios.
3. Estimar el número medio de libros en los hogares donde el padre tiene estudios medios y en los hogares en los que tiene estudios superiores. ¿Puede afirmarse con rotundidad que en este segundo tipo de hogares hay un número de libros superior al de los hogares en los que el padre tiene estudios medios?
4. Estimar la proporción de hogares que cuentan con más de 100 libros.

Solución:

Suponemos que la variable X, *número de libros en un hogar*, sigue una cierta distribución que, en principio, no tenemos por qué determinar, de media μ. El mejor estimador de μ es la media muestral; por tanto,

$$\hat{\mu} = \overline{X} = \frac{87+52+\cdots+5}{20} = 76.2 \quad \text{libros.}$$

De aquí, estimamos que en los hogares de esta ciudad hay 76.2 libros por término medio.

Para estimar la proporción de familias, p, en las que el padre tiene estudios primarios, se utiliza el resultado conocido de que el mejor estimador de p es la proporción muestral. Como la muestra cuenta con 20 familias, teniendo en 6 de ellas el padre estudios primarios, deducimos que

$$\hat{p} = \frac{6}{20} = 0.30,$$

es decir, estimamos que en el 30% de las familias de la ciudad, el padre tiene sólo estudios primarios.

Tenemos que estimar la media, μ_1, del número de libros en los hogares en los que el padre tiene estudios superiores, y la media, μ_2, del número de libros en los hogares en los que el padre tiene estudios medios.

La tabla recoge el número de libros en los distintos tipos de familias:

Familias con padre con E.S.	*Familias con padre con E.M.*
87	62
150	93
104	43
240	52
153	79
	90
	31
	101
	79

Estimamos que en los hogares en los que el padre tiene un mayor nivel de estudios, hay más libros por término medio, pero esta afirmación nunca puede ser rotunda, sino que viene condicionada por la aleatoriedad de la muestra. Para poder afirmarlo con rotundidad, tendríamos que tener recogido el número de libros de todas las fami-

lias en las que el padre tiene estudios superiores y el de todas en las que tiene estudios medios.

De los 20 hogares de la muestra sólo hay 5 que tienen más de 100 libros; por tanto, la proporción muestral de hogares con más de 100 libros es

$$\hat{p} = \frac{5}{20} = 0.25,$$

Es decir, estimamos que el 25% de las familias de la ciudad cuentan con más de 100 libros.

BIBLIOGRAFÍA COMENTADA

PÉREZ y LÓPEZ (PÉREZ, R. y LÓPEZ, A. J. *Análisis de datos económicos II. Métodos inferenciales*. Pirámide. Madrid, 1997), capítulos 5 y 6, proporciona una amena y a la vez rigurosa exposición de los conceptos de muestra y estadístico. Enfoques menos formales se pueden encontrar en MOORE (MOORE, D. S. *Estadística aplicada básica*. Antoni Bosch. Barcelona, 1995), capítulos 3 y 4, o en NEWBOLD (NEWBOLD, P. *Estadística para los Negocios y la Economía*. Prentice Hall. Madrid, 1997), capítulos 6 y 7; este último incluye una numerosa colección de ejemplos y ejercicios propuestos, muchos de ellos con datos reales de diferentes campos de aplicación.

Una visión más formal y completa se puede consultar en RUIZ-MAYA y MARTÍN PLIEGO (MARTÍN PLIEGO, F. J. Y RUIZ MAYA, L. *Estadística I: Probabilidad*. AC. Madrid, 1995), capítulos 1, 2, 3, 4, 5 y 6, que incluye una numerosa colección de ejercicios resueltos. Un buen texto en inglés, con rigor y muchos ejemplos, es ROHATGI (ROHATGI, V. K. *Statistical Inference*. Wiley. New York, 1994), capítulo 10.

5. INTERVALOS DE CONFIANZA

1. CONCEPTO Y CONSTRUCCIÓN

La estimación puntual presenta un gran inconveniente: aun utilizando el mejor estimador, no sólo no acertaremos en la estimación (la posibilidad de acertar es remota), sino que desconoceremos el grado de precisión y fiabilidad de la misma. Así, cuando estimábamos que el 23% de los hogares dispone de lavavajillas, no medimos ni la discrepancia con el verdadero valor del parámetro, ni la probabilidad de equivocarse en menos de una cierta cantidad. La única garantía que podíamos tener acerca de la bondad de la estimación provenía del hecho de que se realizase con el estimador más adecuado. Para evitar esta insuficiencia de la estimación puntual se introducen los intervalos de confianza[25].

25. El concepto de intervalo de confianza fue introducido en 1934 por el estadístico ruso de origen polaco Jerzy Neyman (1894-1981) (véase nuevamente Pearson y Kendall (PEARSON, E.S. y KENDALL, M.G. Opus cit. 1970. Pág. 455). Tras trabajar en Polonia en un instituto de investigación agrícola, al igual que R.A. Fisher, emigró primero a Londres y posteriormente a Estados Unidos, donde fue profesor de la Unversidad de California en Bekerley, universidad en la que fundó el Laboratorio de Estadística que dirigió hasta su muerte. Además de sus aportaciones a los intervalos de confianza y a la teoría de muestras, Neyman, junto con Egon Pearson (1895-1980) (véase la observación anterior), propuso una visión de los contrastes de hipótesis más compleja matemáticamente que la propuesta por Fisher, con el que mantuvo una viva polémica. En las pruebas de significación de Fisher, sólo se atiende a una hipótesis, H_0, y se trata de medir la fuerza de la evidencia muestral contra la misma a partir del p-valor; en cambio, en el planteamiento de Neyman, las dos hipótesis, H_0 y H_1, entran en disputa buscándose la regla que produzca una menor probabilidad de error de tipo II (aceptar H_0 siendo falso) entre todas las pruebas con una probabilidad de error tipo I (rechazar H_0 siendo verdadero) no superior a un valor predeterminado a. La dificultad de evaluar la probabilidad de error de tipo II en muchas situaciones hace que dentro de los contrastes de significación, sea el procedimiento de Fisher el que ha prevalecido mayoritariamente.

Sea una variable aleatoria X cuya distribución depende de un parámetro θ; para obtener información sobre este parámetro, tomamos una muestra aleatoria simple de la variable $(X_1, \ldots, X_n)$. Un intervalo de confianza aleatorio a un nivel de $1-\alpha$ es un conjunto de posibles valores del parámetro, dentro del cual se encuentra el verdadero valor del mismo con una probabilidad de $1-\alpha$. Este conjunto está delimitado por dos estadísticos: el primero de ellos, el extremo inferior del intervalo, es un estimador por defecto del parámetro, mientras que el segundo, el extremo superior del intervalo, es un estimador por exceso del mismo. Cuando la muestra se concreta, el intervalo pasa de ser aleatorio a ser un intervalo en la recta real en el que *confiamos* que esté el verdadero valor del parámetro[26].

De manera más formal, $[T_I(X_1, \ldots, X_n), T_S(X_1, \ldots, X_n)]$ es un intervalo de confianza aleatorio para el parámetro θ a nivel $1-\alpha$ en una muestra aleatoria simple $(X_1, \ldots, X_n)$ si

$$p[T_I(X_1,\ldots X_n) \le \theta \le T_S(X_1,\ldots X_n)] = 1-\alpha.$$

Para la realización de la muestra $(x_1, \ldots, x_n)$,obtenemos el intervalo de confianza numérico

$$[T_I(x_1,\ldots x_n), T_S(x_1,\ldots x_n)].$$

Ejemplo 1. Queremos saber acerca del número de horas diarias de estudio de los bachilleres españoles, para lo cual tomamos una muestra de tamaño 1000 que arroja los resultados que se incluyen en la tabla:

Horas de estudio	3.2	0.8	1.6	...	2.9	2.3	4.1

siendo 2.7 el número medio de horas diarias que dedican al estudio los 1000 bachilleres seleccionados.

26. Un intervalo de confianza puede utilizarse para tomar decisiones sobre el verdadero valor del parámetro. Así, planteada una hipótesis sobre q, H_0: $q=q_0$, se acepta (no se rechaza) si q_0 es uno de los valores del intervalo. Véase el siguiente capítulo, en el que se abordan los contrastes de hipótesis.

La siguiente gráfica muestra el histograma de frecuencias de estos datos:

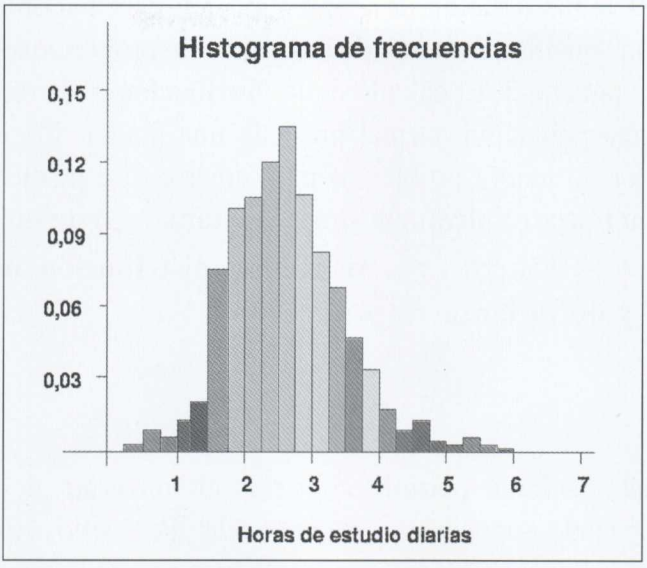

Figura 5.1

Vamos a construir el intervalo de confianza para μ a un nivel de significación de 0.95, esto es, con una probabilidad de equivocarnos de 0.05.

Para abordar el problema suponemos que X, *número de horas de estudio diarias de un bachiller*, sigue una distribución normal de media μ, desconocida, y de varianza 0.81.

La suposición de normalidad está plenamente justificada dada la naturaleza de la variable, que se ve influida por múltiples factores; esta suposición se ve corroborada por la forma que presenta el histograma anterior, que no es muy diferente a la función de densidad de una normal. Por otro lado, la suposición de varianza conocida carece de fundamento (si la media es desconocida, con más motivo lo será también la varianza), pero esta suposición sirve para introducir el problema sin excesivas complicaciones formales.

Por tanto,

$$X \to N(\mu, \sigma = 0.9).$$

La media muestral, el mejor estimador de μ, es de nuevo la clave para encontrar un intervalo de confianza sobre μ a nivel de

confianza de $1-\alpha$. Para construir el intervalo de confianza, necesitamos conocer la distribución del estadístico $\overline{X}$. En el capítulo anterior hallábamos la distribución de este estadístico para variables de Bernoulli en un sencillo problema de inferencia de proporciones para una muestra de tamaño 3. El cálculo de la distribución de la media muestral para una población normal presenta una mayor dificultad. Aun así, el lector no tendrá problemas para creerse que **la media muestral de una muestra aleatoria simple de tamaño n de una variable normal**, $\overline{X} \to N(\mu, \sigma/\sqrt{n})$, **tiene una distribución normal de media** μ **y de varianza** σ^2/n:

$$\overline{X} \to N(\mu, \sigma/\sqrt{n}).$$

Este resultado permite construir un intervalo de confianza sobre μ cuando suponemos σ conocida. Para ello, tipifiquemos previamente el estimador $\overline{X}$, restándole la media y dividiendo por su desviación, trasformación, ésta, que no afecta a su normalidad:

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1).$$

A partir de este resultado, buscamos dos valores λ_1 y λ_2 tales que dejen entre sí una probabilidad de $1-\alpha$ en una distribución normal tipificada:

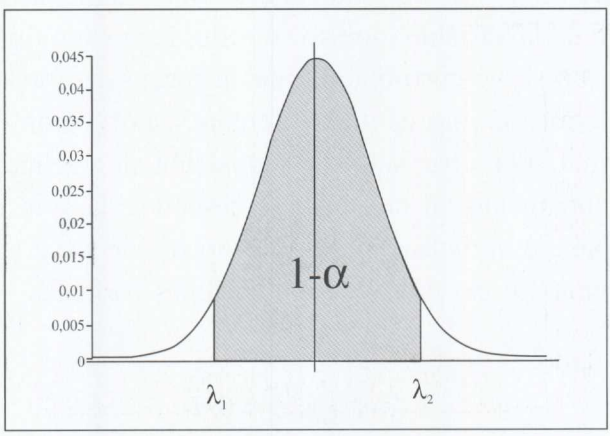

Figura 5.2

De esta forma,

$$p\left[\lambda_1 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq \lambda_2\right] = 1-\alpha.$$

En esta doble desigualdad operamos para dejar solo, y en el centro de las mismas, el parámetro μ sobre el que pretendemos hacer inferencia. De ello resulta

$$p\left[\overline{X}-\lambda_2 \cdot \sigma/\sqrt{n} \leq \mu \leq \overline{X}-\lambda_1 \cdot \sigma/\sqrt{n}\right] = 1-\alpha,$$

que es un intervalo de confianza aleatorio para μ a nivel 1-α.

Notemos que existen muchas parejas (λ_1, λ_2) que dejan entre ellas una probabilidad de 1-α pero, evidentemente, es deseable que la estimación sea lo más precisa posible, esto es, que el intervalo tenga longitud mínima. En este caso, la longitud del intervalo es

$$L = (\overline{X}-\lambda_1 \cdot \sigma/\sqrt{n}) - (\overline{X}-\lambda_2 \cdot \sigma/\sqrt{n}) = (\lambda_2-\lambda_1)\sigma/\sqrt{n},$$

que se hace mínima cuando λ_1 y λ_2 estén lo más cerca posible, situación que se da cuando son simétricos. Entonces, a λ_1 y λ_2 los denotaremos como

$$\lambda_1 = -z_{\alpha/2} \qquad \lambda_2 = z_{\alpha/2}.$$

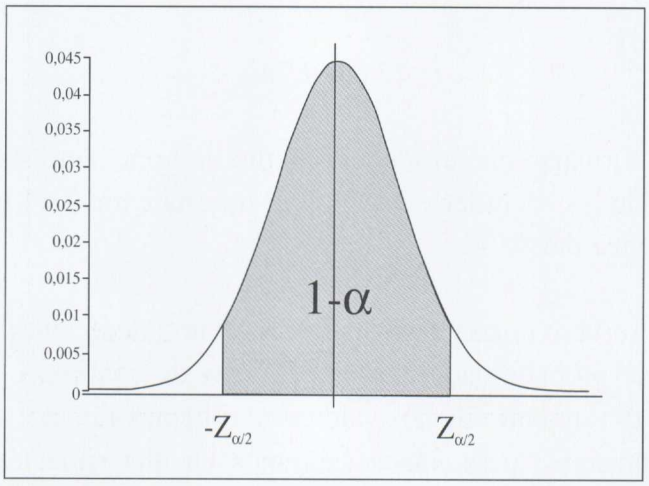

Figura 5.3

De esta forma, el intervalo óptimo (el más corto) es

$$p\left[\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}\right] = 1-\alpha\,.$$

En concreto, para construir un intervalo de confianza al 95% sobre el número medio de horas de estudio, μ, con una muestra de tamaño 1000, sustituyendo se obtiene:

$$p\left[\overline{X} - 1.96 \cdot 0.9/\sqrt{1000} \leq \mu \leq \overline{X} + 1.96 \cdot 0.9/\sqrt{1000}\right] = 0.95\,,$$

donde –1.96 y 1.96 son los dos puntos que en la distribución normal estándar dejan 0.025 de probabilidad en cada cola.

Simplificando resulta

$$p\left[\overline{X} - 0.056 \leq \mu \leq \overline{X} + 0.056\right] = 0.95\,,$$

lo que permite decir que el verdadero valor del parámetro μ está entre $\overline{X} - 0.056$ y $\overline{X} + 0.056$, con una probabilidad de 0.95. Estos dos estadísticos son estimadores por defecto y por exceso, respectivamente, de μ.

Finalmente, dado que el número medio de horas de estudio entre los bachilleres de la muestra era 2.7, se sustituye en la anterior expresión dando lugar al intervalo real

[2.64, 2.75]

pudiendo afirmarse que el número medio de horas diarias dedicadas al estudio de los bachilleres españoles, μ, está entre 2.64 y 2.75 con una confianza del 95%.

Obsérvese que el anterior intervalo no puede interpretarse en términos de probabilidad, sino en términos de confianza. Si hemos acertado, μ está entre dichos valores, y si hemos fallado, μ no está entre los mismos, pero nunca sabremos en qué situación nos encontramos.

Si este problema se plantea repetidas veces tomando cada vez una muestra distinta, obtendríamos intervalos de confianza (no aleatorios) distintos en cada caso, pudiendo afirmarse que en el 95% de esos intervalos, hemos acertado, y que en el 5% restante, hemos fallado (nunca podremos identificar cuáles son aquéllos en los que hemos acertado y aquéllos en los que hemos fallado)

Nótese que en la construcción del intervalo de confianza podemos controlar tres factores:

— El tamaño de la muestra n que nos cuantifica el número de observaciones y, por tanto, la *cantidad de información* de que dispondremos.
— La fiabilidad del intervalo, $1-\alpha$, esto es, la probabilidad de que el parámetro se encuentre dentro del intervalo aleatorio; o lo que es lo mismo, la probabilidad de equivocarse, α.
— La precisión de la estimación o longitud del intervalo.

Así, en el intervalo construido, contábamos con una muestra de tamaño 1000 y para una fiabilidad de 0.95, hemos obtenido un intervalo de longitud 0.12, o en notación más habitual, una precisión de ± 006.

Fijado uno de los tres factores anteriores, podemos ver cómo se relacionan los otros dos entre sí. Desarrollemos estos resultados en el supuesto del ejemplo planteado, si bien todos ellos son generalizables a cualquier otra situación.

I. Así, fijado el tamaño de la muestra, n, mayor fiabilidad (es decir, menor α) implica una menor precisión (un intervalo más largo); esto es, si queremos incrementar la probabilidad de acierto, lo haremos a expensas de perder precisión en la estimación.

La longitud del intervalo de confianza óptimo para la media poblacional, μ, de una variable normal con desviación σ conocida vale:

$$L = (\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}) - (\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}) = 2z_{\alpha/2}\,\sigma/\sqrt{n}\,.$$

Fijado el tamaño n, al crecer α (menor fiabilidad), $z_{\alpha/2}$ decrece, para dejar a su derecha una cola más grande, y L también decrece (mayor precisión). Por tanto, la longitud del intervalo es función decreciente de α. Razonando de la misma forma, vemos que al crecer la longitud (menor precisión), el valor de α decrece (mayor fiabilidad).

Ilustraremos este resultado en el ejemplo propuesto. Como vimos, el intervalo de confianza al 95% sobre el número medio diario de horas de estudio de los bachilleres era

$$[2.64,\ \ 2.75]$$

siendo la precisión de ± 0.06. Construyamos ahora el intervalo con una mayor fiabilidad, en concreto, al 99%. Como vimos, la expresión general del intervalo, esto es, sin concretar el valor de α, era:

$$p\left[\overline{X} - z_{\alpha/2} \cdot 0.9/\sqrt{1000} \le \mu \le \overline{X} + z_{\alpha/2} \cdot 0.9/\sqrt{1000}\right] = 1 - \alpha\,.$$

En este caso, $1-\alpha = 0{,}99$ y, por tanto, $\alpha = 0.01$.

La tabla de la distribución normal estándar nos proporciona el nuevo valor de $z_{\alpha/2}$, que resulta ser 2.57. Sustituyendo,

$$p\left[\overline{X} - 2.57 \cdot 0.9/\sqrt{1000} \le \mu \le \overline{X} + 2.57 \cdot 0.9/\sqrt{1000}\right] = 0.99\,,$$

o, lo que es lo mismo,

$$p\left[\overline{X} - 0.073 \le \mu \le \overline{X} + 0.073\right] = 0.99\,.$$

Concretando el valor de la media muestral para la muestra observada, se obtiene el intervalo numérico

$$[2.63,\ \ 2.77]$$

intervalo de confianza más largo que el obtenido anteriormente, esto es, su precisión, ± 0.073, es sensiblemente menor que la del intervalo anterior.

Todo esto nos confirma la imposibilidad de encontrar un intervalo *ideal*, muy fiable y muy preciso, teniendo que llegar a situaciones de compromiso en las que no se sacrifique la precisión para conseguir una fiabilidad óptima y viceversa.

II. Asimismo, para una fiabilidad concreta, un aumento en el tamaño de la muestra produce una mejora en la precisión de la estimación. Esto es, si α es fijo, al aumentar el tamaño muestral n, la longitud L del intervalo decrece.

De nuevo, si consideramos la expresión de la longitud del intervalo,

$$L = 2z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

vemos que ésta es función decreciente de n. Por tanto, al aumentar n (mayor información), la longitud L decrece (intervalo más preciso).

Volvamos al ejemplo y consideremos que por motivos presupuestarios sólo se ha podido encuestar a 100 bachilleres, siendo en este caso, el número medio de horas diarias de estudio de 2.65. El intervalo de confianza para μ, media poblacional de las horas diarias de estudio, al 95%, para una muestra de tamaño 100 es:

$$p\left[\overline{X} - 1.96 \cdot 0.9/\sqrt{100} \leq \mu \leq \overline{X} + 1.96 \cdot 0.9/\sqrt{100}\right] = 0.95$$

esto es,

$$p\left[\overline{X} - 0.176 \leq \mu \leq \overline{X} + 0.176\right] = 0.95\,.$$

Sustituyendo por la media muestral observada resulta el intervalo numérico

[2.47, 2.82]

cuya precisión ±0.176 es muy inferior a la obtenida con la muestra de tamaño 1000.

III. De la misma forma, para una precisión fijada, un aumento en el tamaño muestral produce una mayor fiabilidad.

De la expresión de la longitud del intervalo deducimos que

$$z_{\alpha/2} = \frac{L \cdot \sqrt{n}}{2\sigma}$$

Entonces, si L permanece fijo, un aumento de n produce un aumento de $z_{\alpha/2}$, o lo que es lo mismo, una disminución en la probabilidad α. Por tanto, si pretendemos que el intervalo tenga una longitud determinada y podemos aumentar el tamaño de la muestra, este aumento provoca una mayor fiabilidad en la estimación (α disminuye)

Estas dos últimas observaciones evidencian un resultado totalmente esperable: la posibilidad de contar con una muestra más grande mejora la estimación bien sea aumentando la fiabilidad (disminuyendo α), bien sea aumentando la precisión (disminuyendo L).

Ahora bien, este deseable aumento de información no siempre es posible. Pensemos que una muestra más grande supone un mayor coste económico, una mayor demora en la obtención de resultados e, incluso, una pérdida en la calidad de la información. En la práctica, el *cliente* que encarga una encuesta a un estadístico le pide que los resultados obtenidos tengan una cierta fiabilidad (un determinado $1-\alpha$) y una cierta precisión (un determinado L); el estadístico diseñará una muestra lo más pequeña posible (esto es, lo más barata, rápida y *buena* posible) para conseguir dichos objetivos (con todo, esta situación no deja de ser ideal, pues en la mayor parte de las situaciones el *cliente* dispondrá de un techo presupuestario, lo que limitará el número de observaciones a realizar). En esta situación, despejando n en la expresión de la longitud del intervalo, se obtiene:

$$n = \frac{4z_{\alpha/2}^2 \sigma^2}{L^2},$$

de donde podremos obtener el valor de n que nos proporcione un intervalo de confianza de una fiabilidad y una precisión determinada.

Así, y volviendo al ejemplo, el Ministerio de Educación se plantea realizar una encuesta entre los bachilleres para saber, entre otras cosas, el número medio diario de horas que éstos dedican al estudio. El Ministerio encarga la muestra a un gabinete estadístico exigiéndole que los resultados obtenidos tengan una fiabilidad del 90% y una precisión de ±0.01. Si el gabinete estadístico supone que el número de horas de estudio de los bachilleres españoles sigue una distribución normal, de media μ desconocida y varianza 0.81 (lo que es mucho suponer), ¿cuántas encuestas deberán realizarse para conseguir la fiabilidad y precisión solicitadas?

Como se quiere una fiabilidad del 90%, es decir, $1-\alpha = 0.90$, buscaremos en las tablas de la normal los valores $z_{\alpha/2}$ y $-z_{\alpha/2}$ que dejen a su derecha y a su izquierda, respectivamente, una probabilidad de 0.05 ($\alpha = 0.05 + 0.05 = 0.10$). Estos valores resultan ser 1.64 y –1.64.

Además, una precisión de ±0.01 supone que el intervalo tenga una longitud de 0.02. Sustituyendo en la expresión que proporciona n a partir de $z_{\alpha/2}$ y de L, obtenemos

$$n = \frac{4(1.64)^2 \cdot 0.81}{0.02^2} = 21785.76$$

Por tanto, hay que realizar un mínimo de 21.786 encuestas para conseguir una fiabilidad del 90% y una precisión de ± 0,01 (un mayor número de encuestas mejorará la fiabilidad y precisión pero encarecerá y demorará la encuesta con el consiguiente perjuicio para el *cliente*).

El lector habrá advertido que la varianza σ^2 juega también un papel importante en la estimación por intervalos. En concreto, las

variables menos dispersas (menos *variables*), es decir, las que tienen varianza pequeña, admiten una mejor estimación, en el sentido de una estimación más fiable y precisa.

El cuadro siguiente resume la construcción de un intervalo de confianza para la media poblacional con α conocido:

$X \to N(\mu, \sigma)$ con σ conocido

Muestra aleatoria simple: $(X_1, \ldots, X_n)$

Estadístico: $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$

Intervalo de confianza para μ a nivel $1-\alpha$:

$$p\left[\overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \le \mu \le \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}\right] = 1-\alpha$$

2. INTERVALOS DE CONFIANZA PARA LA MEDIA DE UNA POBLACIÓN NORMAL

Todos los desarrollos anteriores nos han permitido *entender* la construcción de un intervalo de confianza y las principales virtualidades de este tipo de estimación. Ahora bien, presentan el inconveniente de partir de una suposición irreal: el conocimiento del valor de la varianza σ^2. ¿Qué ocurre ante el planteamiento realista de suponer que σ^2 es desconocido? En esta situación, el tamaño de la muestra va a condicionar la elección del estadístico que permitirá la construcción del intervalo.

Caso 1. Tamaño muestral, n, grande

Si queremos construir un intervalo de confianza sobre la media poblacional μ de una variable normal con σ desconocido y contamos con una muestra grande, utilizaremos el estadístico

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

al igual que hicimos bajo el supuesto irreal de que σ fuese conocido. Ahora bien, como n es grande, estimaremos previamente σ^2, bien sea con la varianza muestral, S_X^2, bien sea con la cuasivarianza, S_C^2, que en este caso prácticamente coincidirán. El tamaño de la muestra nos asegura una buena estimación de σ^2, por lo que

$$\frac{\overline{X}-\mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{aprox.} N(0,1),$$

siendo ésta una buena aproximación.

A partir de este estadístico y, como ya vimos en el caso anterior, construiremos el correspondiente intervalo,

$$p\left[\overline{X}-z_{\alpha/2}\cdot\hat{\sigma}/\sqrt{n}\le\mu\le\overline{X}+z_{\alpha/2}\cdot\hat{\sigma}/\sqrt{n}\right]=1-\alpha$$

intervalo de confianza para μ cuando el tamaño muestral es grande. Pero, ¿a partir de qué valores consideraremos n grande? No existe unanimidad entre los autores a la hora de fijar este convenio; unos hablan de n mayor que 20 o 30, mientras que otros exigen valores superiores a 100.

Caso 2. Tamaño muestral, n, pequeño

Cuando el tamaño muestral es pequeño, la aproximación planteada en el caso anterior es inconsistente, lo cual nos lleva a buscar otro estadístico[27].

27. Históricamente el problema se lo planteó en la década de los veinte el estadístico W.S. Gosset (1876, 1937), más conocido por el seudónimo de *Student* que utilizaba en sus actividades investigadoras. En su trabajo profesional en el departamento de control de calidad de la fábrica de cervezas Guinness, se enfrentaba al problema de estimar y comparar el contenido medio de ciertas sustancias en las cervezas, disponiendo de un número reducido de muestras para analizar. Para solucionar este inconveniente, Gosset estudia la distribución de un nuevo estadístico,

$$\frac{\overline{X}-\mu}{S_c/\sqrt{n}},$$

obteniendo una nueva distribución que, a partir de entonces, lleva su nombre: la distribución t de Student. El resultado obtenido por Gosset se formaliza de la siguiente manera:

La distribución t de Student presenta un perfil similar a una normal estándar, si bien ésta presenta mayor probabilidad en el centro y menor en las colas, como muestra el gráfico:

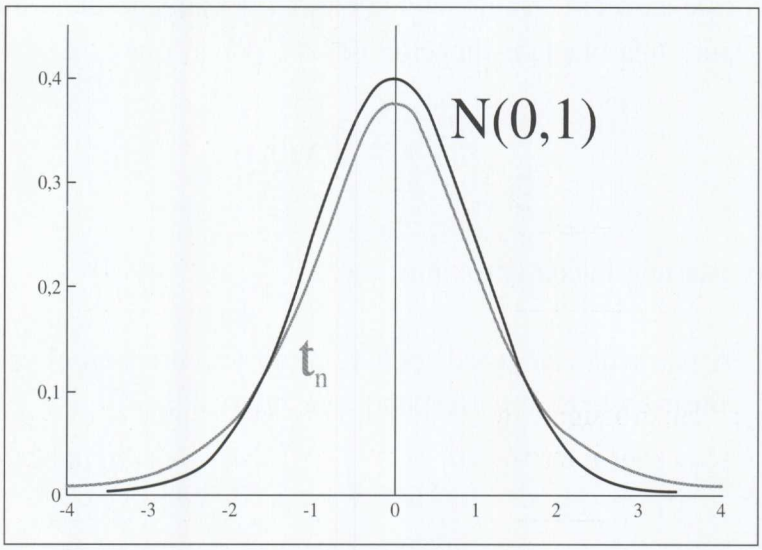

Figura 5.4

Cuando los grados de libertad de la t de Student son muy grandes, esta distribución prácticamente coincide con la normal estándar.

A partir del estadístico de la t de Student, se construye, siguiendo los mismos pasos que para el estadístico de la normal, el correspondiente intervalo aleatorio de confianza:

$$p\left[\overline{X} - t_{\alpha/2} \cdot S_c/\sqrt{n} \le \mu \le \overline{X} + t_{\alpha/2} \cdot S_c/\sqrt{n}\right] = 1-\alpha$$

donde $t_{\alpha/2}$ y $-t_{\alpha/2}$ son los valores simétricos que en la correspondiente distribución t de Student, dejan a la derecha y a la izquierda, respectivamente, una probabilidad de $\alpha/2$.

Si $(X_1, \ldots, X_n)$ es una muestra aleatoria simple de tamaño n de una distribución $N(\mu, \sigma)$, entonces el estadístico

$$\frac{\overline{X} - \mu}{S_c/\sqrt{n}}$$

sigue una distribución t de Student con *n-1* grados de libertad. Esto es,

$$\frac{\overline{X} - \mu}{S_c/\sqrt{n}} \to t_{n-1}.$$

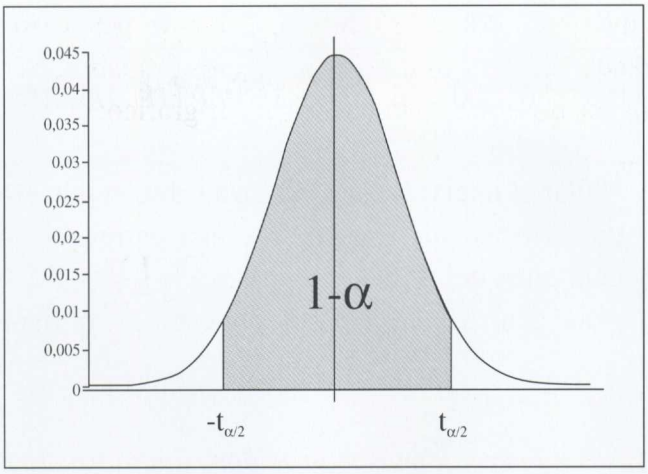

Figura 5.5

El cuadro siguiente resume las dos situaciones en las que nos planteamos hallar intervalos de confianza sobre la media poblacional.

$X \to N(\mu, \sigma)$ con σ desconocido Muestra aleatoria simple: $(X_1, \ldots, X_n)$

I. n grande

Estadístico: $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$, estimando previamente σ.

Intervalo de confianza para μ a nivel $1-\alpha$:

$$p\left[\overline{X} - z_{\alpha/2} \cdot \hat{\sigma}/\sqrt{n} \le \mu \le \overline{X} + z_{\alpha/2} \cdot \hat{\sigma}/\sqrt{n}\right] \approx 1-\alpha$$

I. n pequeño

Estadístico: $\dfrac{\overline{X}-\mu}{S_c/\sqrt{n}} \to t_{n-1}$

Intervalo de confianza para μ a nivel $1-\alpha$:

$$p\left[\overline{X} - t_{\alpha/2} \cdot S_c/\sqrt{n} \le \mu \le \overline{X} + t_{\alpha/2} \cdot S_c/\sqrt{n}\right] = 1-\alpha$$

Ejemplo 2. El profesor de gimnasia del instituto está interesado en conocer cómo afecta el esfuerzo en la carrera a las pulsaciones de los alumnos. Para ello, eligió al azar a 12 alumnos, les sometió al "Test de Cooper" (12 minutos de carrera continuada) y les tomó sus pulsaciones inmediatamente después. Los resultados de esta medición los encontramos en la tabla siguiente:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12
Pulsaciones por minuto	160	186	200	176	208	180	170	168	180	165	160	148

(a) Hallar el intervalo de confianza sobre el número medio de pulsaciones por minuto tras una carrera a un nivel de confianza del 95%.

(b) ¿Qué ocurrirá si queremos aumentar la confianza al 99%?

Solución:

(a) El número de pulsaciones por minuto tras la carrera, X, es una variable que, como la mayoría de los fenómenos biológicos, se ajusta muy bien a una distribución normal. Por tanto,

$$X \to N(\mu, \sigma) \ .$$

Para hallar el intervalo de confianza sobre μ, como la muestra es pequeña, tendremos que utilizar el estadístico de la t de Student, que, en este caso, se concreta en

$$\frac{\overline{X} - \mu}{S_c / \sqrt{12}} \to t_{11} \ .$$

Entonces, buscaremos en las tablas de la t de Student con 11 grados de libertad, el par de valores simétricos que dejan entre sí una probabilidad de 0.95, siendo estos valores –2.20 y +2.20.

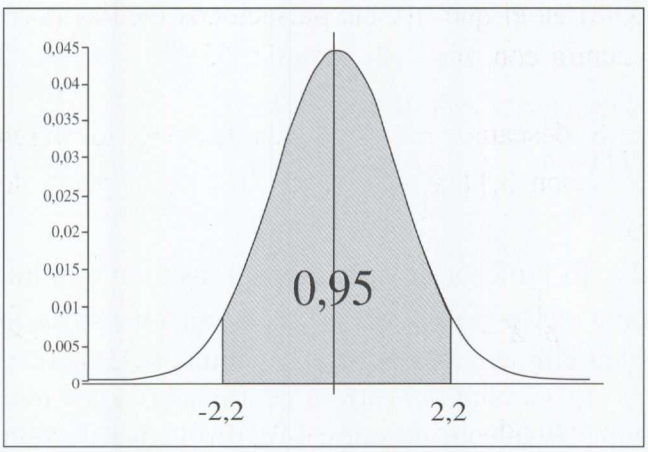

Figura 5.6

Por tanto,

$$p\left[-2.20 \leq \frac{\overline{X}-\mu}{S_c/\sqrt{12}} \leq 2.20\right] = 0.95$$

Operando, dejaremos el parámetro μ solo y en el centro de la desigualdad:

$$p\left[\overline{X} - 2.20\frac{S_c}{\sqrt{12}} \leq \mu \leq \overline{X} + 2.20\frac{S_c}{\sqrt{12}}\right] = 0.95\,.$$

Así,

$$\left[\overline{X} - 2.20\frac{S_c}{\sqrt{12}},\quad \overline{X} + 2.20\frac{S_c}{\sqrt{12}}\right]$$

es el intervalo aleatorio en el que podemos afirmar que se encuentra μ con una probabilidad de 0.95. Calculemos la varianza y la cuasivarianza de la muestra:

$$\overline{X} = 175.08$$

y

$$S_c^2 = 294.45$$

Sustituyendo, encontramos el intervalo numérico [164,18, 185,98] en el que μ, las pulsaciones medias por minuto, se encuentra con una confianza del 95%.

(b) Si deseamos una confianza del 99%, los valores $t_{\alpha/2}$ y $-t_{\alpha/2}$ valen 3.11 y –3.11, con lo cual el intervalo aleatorio será:

$$p\left[\overline{X} - 3.11\frac{S_c}{\sqrt{12}} \leq \mu \leq \overline{X} + 3.11\frac{S_c}{\sqrt{12}}\right] = 0.99\,,$$

y, concretando los valores muestrales de la media y la cuasidesviación, se obtiene el intervalo numérico [159,67,

190,49]. Como se ve, el intervalo hallado en (a) tenía una precisión de ±10,9 para una confianza del 95%; el nuevo intervalo ha perdido precisión, ahora es de ±15,41, al aumentar la fiabilidad al 99%.

Ejemplo 3. Un fabricante de bolígrafos quiere estimar su duración media, medida en kilómetros de escritura. Para ello, el laboratorio de control de calidad ha hecho escribir a 150 bolígrafos sobre un mecanismo de papel continuo, midiendo los kilómetros de escritura hasta su agotamiento. Para esta muestra ha obtenido una duración media de 6.3 kilómetros con una desviación muestral de 0.7 kilómetros.

(a) Hallar un intervalo de confianza al 95% para la duración media de los bolígrafos.
(b) Si el laboratorio pretendiese obtener una mayor fiabilidad, 99%, y una precisión de ±0.05 ¿Cuántos bolígrafos se hubiesen necesitado para la prueba?

Solución

(a) Supongamos que *la duración, en kilómetros de escritura, de un bolígrafo*, X, sigue una distribución normal:

$$X \to N(\mu, \sigma)$$

Dado que tenemos una muestra suficientemente grande, para construir un intervalo de confianza sobre μ, duración media, podemos utilizar el estadístico de la normal:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1) .$$

Ahora bien, como σ es desconocido, lo estimaremos previamente, bien sea con la desviación muestral, bien con la cuasidesviación. En el primer caso:

$$\hat{\sigma} = S_X = 0.7 .$$

Si utilizásemos la cuasidesviación y dado que

$$n \cdot S_X^2 = (n-1) \cdot S_c^2,$$

entonces,

$$\hat{\sigma} = S_c = \sqrt{\frac{n}{n-1}}\, S_X = 0.702345,$$

valor que difiere muy poco del obtenido con la desviación muestral. Utilizando, por comodidad, la primera de las estimaciones, resulta:

$$\frac{\overline{X} - \mu}{0.7/\sqrt{150}} \xrightarrow{aprox.} N(0,1).$$

Hallando los correspondientes valores de $-z_{\alpha/2}$ y $z_{\alpha/2}$ en la tabla de la normal estándar:

$$p\left[-1.96 \le \frac{\overline{X} - \mu}{0.7/\sqrt{150}} \le 1.96\right] \approx 0.95,$$

de donde, despejando μ, obtenemos el intervalo aleatorio:

$$p\left[\overline{X} - 0.112 \le \mu \le \overline{X} + 0.112\right] \approx 0.95.$$

Sustituyendo por el valor observado de la media muestral, hallamos el intervalo numérico:

[6.19, 6.41]

en el que se encuentra la duración media de los bolígrafos, μ, con una confianza del 95%.

(b) Para encontrar el tamaño muestral necesario para conseguir una fiabilidad del 99% y una precisión de ±0,05, recurriremos a la expresión deducida anteriormente:

$$n = \frac{4z_{\alpha/2}^2 \cdot \sigma^2}{L^2}.$$

En este caso, dado que $1-\alpha = 0.99$, el valor de $z_{\alpha/2}$ obtenido en las tablas es 2.57; asimismo, la precisión de ±0,05 equivale a una longitud de 0.1. Como se observará, necesitamos conocer σ (situación irreal) o un estimador suyo (en principio, antes de tomar la muestra). En este caso nos valdremos de la estimación de σ que nos proporciona el apartado (a), esto es, $\hat{\sigma} = 0.7$. Entonces,

$$n = \frac{4 \cdot 2.57^2 \cdot 0.7^2}{0.1^2} = 1294.6,$$

es decir, habría que probar al menos 1295 bolígrafos para obtener la fiabilidad y precisión deseadas.

3. INTERVALOS DE CONFIANZA PARA UNA PROPORCIÓN

Todos los intervalos de confianza anteriores se construyen bajo el supuesto de que la variable sigue una distribución normal. ¿Qué ocurre si la distribución es otra? En base al denominado **Teorema del Límite Central** que afirma que **bajo condiciones muy generales, la media muestral de una muestra aleatoria simple de una variable no normal se distribuye aproximadamente como una normal**, en concreto,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{aprox.} N(0,1)$$

podemos generalizar los intervalos de confianza anteriores para la media de poblaciones no normales siempre que contemos con una muestra suficientemente grande.

X variable no normal de media μ y varianza σ^2
Muestra aleatoria simple: $(X_1, \ldots, X_n)$ con n grande.

Estadístico: $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{aprox.} N(0,1)$,

Intervalo de confianza aproximado para μ a nivel 1-α

$$p\left[\overline{X}-z_{\alpha/2}\cdot\hat{\sigma}/\sqrt{n}\le\mu\le\overline{X}+z_{\alpha/2}\cdot\hat{\sigma}/\sqrt{n}\right]\approx 1-\alpha\,.$$

Un caso especial de estos intervalos de confianza aproximados los encontramos al hacer inferencia sobre una proporción.

Ejemplo 4. Nos planteamos hallar un intervalo de confianza al 95% sobre el porcentaje de españoles partidarios de implantar la pena de muerte para delitos especiales. Para decidir al respecto, se encuestó a 2.000 personas, de las cuales, 435 se mostraron favorables a dicha medida.

Podemos modelizar el problema mediante una variable de Bernoulli:

$$X=\begin{cases}1 & \text{A favor de la pena de muerte} \quad p[X=1]=p\\ 0 & \text{En otro caso} \quad p[X=0]=1-p\end{cases}$$

siendo p, la proporción de españoles favorables a dicha medida, el parámetro a estimar. De esta variable hemos tomado una muestra (X_1, X_2,..., X_{2000}), de tamaño 2000.

El denominado Teorema del Límite Central se puede aplicar a variables de Bernoulli, $X\to b(p)$, sin más que recordar que su media es p y su varianza $p(1-p)$. Entonces,

$$\frac{\overline{X}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \xrightarrow{aprox.} N(0,1)\,,$$

aproximación que es aceptable para muestras grandes[28]. Utilizando este estadístico podemos construir el intervalo de confianza para p. Así, buscaremos los valores $-z_{\alpha/2}$ y $z_{\alpha/2}$ tal que,

28. De hecho, esta fue la primera formulación histórica del Teorema del Límite Central realizada en 1733 por A. De Moivre (1667-1754) (remitimos al lector nuevamente a las referencias Stigler (STIGLER, S.M. Opus cit. 1986. Pág. 70) o Daw y Pearson (DAW, R. H. y PEARSON, E.S. "*Studies in the history of probability and statistics. XXX : Abraham de Moivre's 1733 derivation of the normal curve : a bibliographical note.*" *Biométrica*, 59. 1972).

$$p\left[-z_{\alpha/2} \leq \frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right] \cong 1-\alpha$$

Como es habitual, dejaremos solo y en el centro de la doble desigualdad al valor del parámetro:

$$p\left[\overline{X}-z_{\alpha/2}\cdot\sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{X}+z_{\alpha/2}\cdot\sqrt{\frac{p(1-p)}{n}}\right] \cong 1-\alpha\,.$$

Obsérvese que este intervalo presenta una peculiaridad: el parámetro a estimar aparece en los extremos del intervalo. Para evitar esto, sustituiremos esos valores de p por su estimación, la media muestral, es decir, $\hat{p}=\overline{X}$, lo que provoca que este intervalo sea aproximado en un doble sentido: por la distribución del estadístico y por la sustitución del parámetro por su estimador. El intervalo de confianza para una proporción p cuando n es grande resulta:

$$p\left[\overline{X}-z_{\alpha/2}\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq p \leq \overline{X}+z_{\alpha/2}\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right] \cong 1-\alpha\,.$$

Entonces, para estimar por intervalos el porcentaje de españoles favorables a la pena de muerte, tendremos en cuenta que la proporción muestral vale 0.2175,

$$\hat{p}=\frac{435}{2000}=0.2175\,,$$

y que el valor de $z_{\alpha/2}$ es 1.96. Así, el intervalo aleatorio será

$$p\left[\overline{X}-1.96\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{2000}} \leq p \leq \overline{X}+1.96\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{2000}}\right] \cong 0.95$$

y el intervalo numérico [0.1994, 0.2356]. Es decir, el porcentaje de españoles partidarios de la pena de muerte está entre el 19.94% y el 23.56% con una confianza del 95%. La precisión del intervalo se acostumbra a dar en porcentaje de las desviaciones al centro; en este caso, la precisión será de ±1,81.

Como en el caso de la estimación de la media en una población normal, la forma natural de plantearse la estimación por intervalo de una proporción es exigir una cierta fiabilidad y precisión, buscando el tamaño muestral necesario para conseguirlas. La longitud del intervalo de confianza para p resulta:

$$L = \left(\overline{X} + z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \right) - \left(\overline{X} - z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \right) =$$

$$= 2z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

De aquí, podremos calcular el valor de n en función de la longitud del intervalo, L, y de su fiabilidad, 1-α:

$$n = \frac{4z_{\alpha/2}^2 \cdot \overline{X}(1-\overline{X})}{L^2}.$$

Adviértase que llegamos a un resultado en principio incongruente: queremos saber cuántas observaciones tenemos que realizar y para ello necesitaremos conocer una característica muestral $\overline{X}$, que, obviamente, sólo conoceremos una vez hayamos realizado las observaciones. ¿Cómo solucionar este problema? Existen tres posibles vías:

a) Si tuviésemos información (encuestas anteriores, opiniones de experto,...) sobre el posible valor de la proporción a estimar, sustituiríamos este valor en la anterior expresión.

b) Podríamos realizar una pequeña encuesta (**encuesta piloto**) que nos proporcionase una primera evaluación de la

proporción muestral, $\overline{X}$. Además, esta encuesta puede servir para probar y reformar el cuestionario, organizar el trabajo de campo, etc.

c) Si no contásemos con información alguna ni tuviésemos la posibilidad de realizar la encuesta piloto, nos pondríamos en la situación más desfavorable, esto es, la que da lugar al tamaño muestral más grande para la fiabilidad y precisión deseadas. Esa situación se produce cuando $\overline{X}(1-\overline{X})$ alcanza su máximo, lo cual ocurre cuando $\overline{X} = 0.5$.

El siguiente cuadro resume la construcción de intervalos de confianza para proporciones:

$X \to b(p)$, donde p es la proporción, desconocida, de una cierta característica de la población

Muestra aleatoria simple: $(X_1, \ldots, X_n)$

I. *n* grande

Estadístico: $$\frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \xrightarrow{aprox.} N(0,1) ,$$

Intervalo de confianza para p a nivel $1-\alpha$:

$$p\left[\overline{X} - z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \le p \le \overline{X} + z_{\alpha/2} \cdot \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right] \cong 1-\alpha$$

I. *n* pequeño

Existen tablas y gráficos que proporcionan intervalos para ciertos valores de n y de la proporción muestral, $\hat{p} = \overline{X}$.

Ejemplo 5. Se quiere realizar un estudio sobre el consumo de bebidas alcohólicas entre los bachilleres españoles, para lo cual se realiza una encuesta. De los 3.500 estudiantes encuestados, 893 admitieron consumir algún tipo de bebida alcohólica diariamente.

(a) Hallar un intervalo de confianza al 90% sobre la proporción de bachilleres que beben a diario.

(b) ¿Cuántas encuestas habría que realizar si queremos tener una fiabilidad del 99% y una precisión de ±0.5%?

Solución

(a) Dado que el tamaño de la muestra es grande, 3.500, podemos aplicar el resultado aproximado que proporciona el Teorema del Límite Central para variables de Bernoulli:

$$\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{aprox.} N(0,1)$$

donde los dos valores de p que aparecen en el denominador de la expresión se estiman previamente con la media muestral. Por tanto, siendo –1.64 y 1.64 los valores que dejan entre sí una probabilidad de 0.90 en una normal estándar, resulta:

$$p\left[-1.64 \le \frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{3500}}} \le 1.64\right] \cong 0.90\,.$$

Despejando en el centro de la doble desigualdad, obtenemos

$$p\left[\overline{X}-1.64\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{3500}} \le p \le \overline{X}+1.64\cdot\sqrt{\frac{\overline{X}(1-\overline{X})}{3500}}\right] \cong 0.90\,.$$

Sustituyendo por la proporción muestral observada,

$$\hat{p} = \overline{X} = \frac{893}{3500} = 0.2551\,,$$

hallamos el intervalo numérico [0.2430, 0.2672] en el que p está con una confianza del 90%. Esto es, el porcentaje de españoles que está a favor de la pena capital está entre el 24.30% y el 26.72%, afirmación que mantenemos con una confianza del 90%.

(b) Tenemos que determinar el tamaño muestral necesario para obtener un intervalo de confianza al 99% con una precisión de ±0.5%. Recordemos la expresión, deducida anteriormente, que proporcionaba n a partir de α y L:

$$n = \frac{4z_{\alpha/2}^2 \cdot \bar{X}(1-\bar{X})}{L^2}.$$

En este caso, como $\alpha = 0.10$, la tabla de la distribución normal estándar nos da el valor $z_{\alpha/2} = 2.57$. Por su parte, una precisión de ±0,5% equivale a una longitud de 1% o, escrito en proporción, de 0.01. El valor de la media muestral es, obviamente, desconocido pero podemos utilizar como encuesta piloto la realizada en el apartado (a), esto es, considerar $\bar{X} = 0.2551$ como una primera estimación de p. Entonces,

$$n = \frac{4 \cdot 2.57^2 \cdot 0.25552(1-0.2551)}{(0.01)^2} = 50203.6,$$

es decir, habría que realizar 50.204 encuestas para obtener la precisión y fiabilidad exigidas.

4. OTROS INTERVALOS DE CONFIANZA

En este capítulo hemos desarrollado intervalos sobre la característica de una población (en concreto, sobre la media μ y la proporción p) sin más que conocer la distribución del estadístico adecuado (en concreto, estadísticos funciones de la media muestral, $\bar{X}$). De forma similar podríamos construir intervalos sobre la varianza poblacional[29], σ^2.

29. El intervalo de confianza sobre la varianza de una población normal se construye a partir del estadístico

$$\frac{nS_X^2}{\sigma^2} = \frac{(n-1)S_C^2}{\sigma^2} \to \chi_{n-1}^2,$$

estadístico cuya distribución en una c^2 (ji-cuadrado) con n-1 grados de libertad. Este resultado fue obtenido en 1875 por F.R. Helmert (1843-1917) que en este sentido podría considerarse el "padre" de la nueva distribución. En 1900, Karl Pearson (1857-1936)

Mayor interés, aunque desbordan los límites que nos hemos planteado en este texto, tienen los intervalos cuya finalidad es comparar las características de dos poblaciones:

— Intervalos de confianza para la diferencia de medias, $\mu_1 - \mu_2$, de dos poblaciones, construidos a partir de estadísticos en los que aparecen las diferencias de medias muestrales[30], $\overline{X}_1 - \overline{X}_2$.

— Intervalos de confianza para la diferencia de proporciones, p_1-p_2, una generalización de los anteriores con distribuciones aproximadas[31].

Asimismo, se pueden comparar las varianzas de dos poblaciones construyendo intervalos de confianza para su cociente, σ_1^2 / σ_2^2,

---

utiliza dicha distribución en su contraste de bondad de ajuste; la mayor relevancia de este autor ha hecho que la distribución lleve su nombre. Una descripción de este intervalo se puede ver en Ruiz-Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. J. *Estadística II: Inferencia*, AC. Madrid, 1995. Págs. 265-267).

30. En este caso existen tres estadísticos que permiten construir diferentes intervalos según se verifiquen unas u otras condiciones:

— Un estadístico con distribución normal en el supuesto (irreal) en que las varianzas σ_1^2 y σ_2^2 sean conocidas. El intervalo resultante se puede aplicar en el caso en que las muestras sean grandes.

— Un estadístico con distribución t de Student en el supuesto de que las varianzas desconocidas se puedan considerar iguales.

— Un estadístico con distribución aproximada t de Student cuando no se pueda suponer igualdad de varianzas.

Para una descripción de los tres tipos de intervalos véase Ruiz-Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. J. Opus cit. 1995. Págs. 267-272).

31. El intervalo de confianza sobre el cociente de varianzas de dos poblaciones normales se construye a partir del estadístico

$$\frac{\dfrac{n_1 S_{x_1}^2}{\sigma_1^2}}{\dfrac{n_2 S_{x_2}^2}{\sigma_2^2}} = \frac{\dfrac{(n_1-1)S_{c_1}^2}{\sigma_1^2}}{\dfrac{(n_2-1)S_{c_2}^2}{\sigma_2^2}} \rightarrow F_{n_1-1,n_2-1},$$

cuya distribución es una F de Snedecor con n_1-1 y n_2-1 grados de libertad. Esta nueva distribución es intuida por R A. Fisher (1890-1962) cuando se plantea la distribución del cociente de varianzas muestrales; los estudios fueron concluidos por G. W. Snedecor (1881-1974) que denominó F a dicha distribución en honor a Fisher. En Ruiz-Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. J. Opus cit. 1995. Págs. 273-275), puede verse el desarrollo de este intervalo.

aunque éstos sólo tienen un interés instrumental[32]. En todo caso, y como hemos advertido más arriba, omitiremos estos desarrollos, aunque conocidos los estadísticos correspondientes y sus distribuciones, no presentan ninguna dificultad añadida.

5. EJERCICIOS PROPUESTOS

5.1. *Una muestra de tamaño 16 de un cierto tipo de pilas eléctricas ha presentado una duración media de 225 horas. Suponiendo que la duración de las pilas sigue una distribución normal con* $\sigma = 12$ *horas, hallar un intervalo de confianza al 95% para la duración media poblacional. Si quisiéramos obtener una mayor fiabilidad en la estimación, en concreto un 99%, ¿qué ocurrirá con el nuevo intervalo de confianza?*

5.2. *Una oficina de empleo está interesada en conocer el tiempo en días que un individuo tarda en encontrar trabajo. Para ello se extraen aleatoriamente de su base de datos las fichas correspondientes a 8 individuos recientemente contratados obteniendo los siguientes resultados*:

180 35 155 200 160 120 135 95

Suponiendo que la variable sigue una distribución normal:

a) Estimar puntualmente la media y la varianza.
b) Calcular el porcentaje de individuos que emplean más de 175 días en la búsqueda de trabajo.
c) Hallar un intervalo de confianza al 90 % del tiempo esperado para encontrar empleo.

5.3. *Para conocer la opinión de los alumnos del Instituto sobre la nueva dirección del centro se encuestaron a 150 de ellos, obteniéndose los siguientes resultados:*

32. Véase NEWBOLD, P. *Estadística para los Negocios y la Economía*. Prentice Hall. Madrid, 1997. Págs. 268-270.

— *90 a favor*
— *45 en contra*
— *15 no sabe/no contesta*

a) *Hallar un intervalo al 95% sobre el porcentaje de alumnos que rechaza la nueva dirección.*
b) *En base a los resultados anteriores, calcular el número mínimo de encuestas que habría que realizar para obtener una fiabilidad del 99% y una precisión de ±1.5%.*

5.4. *El profesor de Ciencias Sociales ha propuesto la realización de un trabajo para conocer la situación económica de las familias del barrio. Con esta finalidad se ha encuestado de forma aleatoria a 280 familias obteniéndose, entre otros, los siguientes resultados relativos a sus ingresos mensuales:*

5.5.

Ingresos en euros	*Nº de familias*
480-720	35
720-900	90
900-1200	105
1200-1800	50

a) *Hallar un intervalo de confianza al 90 % sobre los ingresos medios de las familias del barrio.*
b) *Obtener un intervalo de confianza al 95 % sobre la proporción de familias cuyos ingresos superan los 1200 euros.*

BIBLIOGRAFÍA COMENTADA

NEWBOLD (NEWBOLD, P. *Estadística para los Negocios y la Economía*. Prentice Hall. Madrid, 1997), capítulo 8, recorre la mayor parte de los intervalos de interés con amenidad, claridad expositiva y rigor; además, incorpora una amplísima colección de ejercicios, muchos de ellos con datos reales. En PEÑA (PEÑA, D. *Fundamentos de Estadística.* Alianza Editorial. Madrid, 2001), capítulo 8, además de desarrollar con concisión la mayor parte de los intervalos, se exponen dos métodos novedosos de construcción: el muestreo *bootstrap* y el *jackknife*.

Para una exposición con una mayor formalización matemática se puede consultar RUIZ-MAYA y MARTÍN PLIEGO (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. J. *Estadística II: Inferencia*, AC. Madrid, 1995), capítulo 7, o el texto en inglés de ROHATGI (ROHATGI, V. K. *Statistical Inference*. Wiley. New York, 1994), capítulo 10.

6. TEST DE CONTRASTE DE HIPÓTESIS[33]

1. UN MANUAL PARA EL PROFESOR

1.1. Definiciones y criterios

Cuando se realizan experimentos aleatorios o, en concreto, se estudian poblaciones, se emiten hipótesis sobre el comportamiento probabilístico de esos experimentos. Por ejemplo, se dice que una moneda no está cargada, o que las alturas de los alumnos del Instituto siguen una distribución $N(1{,}70,\ 0{,}1)$, o que la muestra que hemos tomado para hacer una estimación es representativa de la población.

Una hipótesis estadística es una afirmación sobre el comportamiento de un experimento aleatorio o de una población.

Como dijimos al hablar de la estimación, muchos de los problemas de la inferencia estadística pueden abordarse suponiendo que la característica que se quiere estudiar de la población sigue una

33. La teoría de los contrastes (esto es, de las contrastaciones) de hipótesis es relativamente reciente. Las definiciones básicas y los métodos generales fueron elaborados por Sir R.A. Fisher, J. Neyman y E.S. Pearson entre 1920 y 1933. No debe confundirse Egon Sharpe Pearson con Karl Pearson, de quien hemos hablado al estudiar los conceptos de correlación lineal, y que era padre del primero. Neyman y Pearson trabajaron juntos en el estudio de los contrastes, mientras que Fisher, enfrentado con Pearson por la sucesión en la cátedra de su padre (que finalmente fue dividida en dos) trabajó separadamente de los primeros. En *MacTutor History of Mathematics* (dentro de la Web de la Escuela de Matemáticas y Estadística de la universidad escocesa de St. Andrews) pueden encontrarse biografías comentadas de estos autores clásicos. En Ruiz Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. Opus cit. 1995. Págs. 340 y 341 o en Peña (PEÑA, D. Opus cit. 2001. Págs. 377 a 382) pueden encontrarse también comentarios históricos sobre el origen de la teoría de los contrastes de hipótesis estadísticas.

distribución de probabilidades que es de un tipo conocido, de forma que las hipótesis que se emiten versarán sobre el parámetro o los parámetros que determinan dicha distribución. Diremos entonces que se trata de hipótesis paramétricas [34]. En este capítulo únicamente abordaremos los contrastes de este tipo de hipótesis.

Ejemplo 1. Se desea estudiar la proporción de alumnos del Instituto que han ido al cine la pasada semana. Esta característica puede describirse mediante una variable de Bernoulli, X, que valdrá la unidad si un alumno elegido al azar ha ido al cine, y cero si no ha ido. El parámetro $p = \Pr[X = 1]$, que determina la distribución de X, es la proporción poblacional. Se emite la hipótesis de que dicha proporción no rebasa el 25%, esto es, que $p \leq 0{,}25$.

Ejemplo 2. Se sospecha que los pupitres utilizados por los alumnos de Bachillerato son muy pequeños. Un alumno tendrá serios problemas para acoplarse en el pupitre si su altura es al menos 1,70. Se sabe (se acepta) que las alturas de los alumnos de Bachillerato siguen una distribución normal, $N(\mu,\sigma)$, y se quiere contrastar la hipótesis de que la altura media supera esa cifra, esto es, que $\mu \geq 1{,}70$.

Hay dos formas de analizar la veracidad de las hipótesis estadísticas que se emiten. Si es posible hacer un censo, se podrá conocer el conjunto de la población y calcular el valor del parámetro. Por ejemplo, si se pregunta a cada alumno del Instituto si ha ido al cine la pasada semana (y nos fiamos de las respuestas obtenidas) conoceremos la proporción, p, que se describe en el ejemplo número 1, y sabremos si $p \leq 0{,}25$ o, por el contrario, $p > 0{,}25$.

---

34. Los contrastes denominados no paramétricos no se estudian en este libro por no formar parte de los currículos correspondientes. Sin embargo son muy ilustrativos sobre la técnica general de la contrastación de hipótesis, y en los estudios estadísticos aplicados conviven, e incluso preceden a los paramétricos. Hipótesis como que una muestra es realmente aleatoria (elegida al azar de la población) o que la altura de los adolescentes sigue una distribución normal, forman parte de este grupo. En Ruiz Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. Opus cit. 1995. Cap. 12 y 13) puede encontrarse una descripción de los más usuales, aunque se trata de una presentación algo farragosa. Más interesante es la descripción de Rohatgi (ROHATGI, V. K. *Statistical Inference*. Wiley. New York, 1994) para los lectores con conocimientos de inglés.

Si sólo es posible elegir una muestra representativa de la población, la hipótesis no puede comprobarse. Se acude entonces a su contrastación con los datos que proporcione dicha muestra.

Un contraste de una hipótesis estadística con los datos de una muestra es un estudio del grado de coherencia, o adecuación, entre los datos y la hipótesis.

Ejemplo 1 *(continuación).* No pudiéndose preguntar a todos los alumnos del Instituto, se elige una muestra de cinco alumnos. Sus respuestas han sido *SNSSN*, donde la *S* indica una respuesta afirmativa (ha ido al cine la semana pasada), y *N* negativa (no ha ido). ¿Cómo puede contrastarse la hipótesis de que $p \leq 0{,}25$?, esto es, ¿cómo puede estudiarse si los datos y la hipótesis son coherentes entre sí?

Una posible estrategia consiste en estimar a partir de la muestra la proporción, p, de alumnos que han ido al cine (proporción poblacional). Como sabemos, una estimación razonable es la proporción muestral, $\hat{p} = \bar{x}$, donde $\bar{x} = \left(\sum_{i=1}^{5} x_i\right)/5$, y donde cada x_i vale 1 ó 0 según que se haya obtenido una respuesta afirmativa o negativa en la encuesta. En nuestro caso,

$$\hat{p} = \frac{1+0+1+1+0}{5} = \frac{3}{5},$$

ya que el resultado muestral (*SNSSN*) puede describirse como

$$(x_1, x_2, x_3, x_4, x_5) = (1,0,1,1,0)$$

En consecuencia, puede solventarse la contrastación de la hipótesis $p \leq 0{,}25$ mediante el siguiente argumento: como la muestra es representativa de la población, $\hat{p} = 3/5$ será una estimación razonable de p. Pero $\hat{p} = 3/5 > 0{,}25$, luego rechazaremos la hipótesis.

La conclusión sería que los datos (*SNSSN*) serían incoherentes con la hipótesis $p \leq 0{,}25$, ya que la estimación de p que se obtiene a partir de aquéllos no verifica ésta.

Ejemplo 2 (*continuación*). En este caso, elegimos una muestra representativa de tamaño 3, que nos ha proporcionado las alturas 1,75, 1,80 y 1,70, esto es, $(x_1, x_2, x_3) = (1{,}75, 1{,}80, 1{,}70)$.

Nuevamente, la media muestral, $\bar{x} = (x_1 + x_2 + x_3)/3 = 1{,}75$ es una estimación razonable de μ, y con el criterio adoptado en el ejemplo anterior aceptaríamos la hipótesis $\mu \geq 1{,}70$ ya que la estimación $\hat{\mu} = 1{,}75$ la verifica.

Obsérvese que las afirmaciones que hemos realizado en los ejemplos anteriores concretan, para los datos obtenidos, un criterio general que puede describirse aun sin haber realizado la encuesta. El criterio, para el primer ejemplo, es el siguiente:

Dado un resultado muestral, $(x_1, x_2, x_3, x_4, x_5)$***, aceptamos la hipótesis si*** $\bar{x} = \dfrac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$ ***es menor o igual que 0,25, y la rechazamos si es mayor.***

Dicho criterio puede describirse de forma abreviada clasificando los posibles resultados muestrales,

$$\aleph = \{(x_1, x_2, x_3, x_4, x_5) \text{ tales que } x_i = 0,1,\ i = 1,...,5\}$$

en dos subconjuntos; uno de ellos,

$$A = \{(x_1, x_2, x_3, x_4, x_5) \text{ tales que } \bar{x} \leq 0{,}25\}$$

recoge los resultados para los que se acepta la hipótesis, y el otro,

$$C = \{(x_1, x_2, x_3, x_4, x_5) \text{ tales que } \bar{x} > 0{,}25\}$$

agrupa aquéllos para los que se rechaza. Es evidente que, con el criterio anterior, siempre tomamos una decisión, esto es, $AUC = \aleph$ y que, como parece natural, nunca aceptamos y rechazamos la hipótesis ($A \cap C = \varnothing$). Esta clasificación recibe el nombre de test de contraste de hipótesis (test para el contraste de hipótesis) o, más sucintamente, se denomina test.

Fijemos las definiciones y las notaciones[35]. En todo lo que sigue supondremos que X es una variable aleatoria que mide una característica numérica de la población, y que θ es un parámetro (desconocido) que mide alguna cantidad relacionada con la distribución de probabilidades de X. Supondremos además que se conoce el campo de variación de θ, que será un subconjunto E de la recta real, por lo que una hipótesis, H, afirmación sobre el comportamiento de θ, podrá describirse mediante la afirmación de que θ pertenece a un subconjunto E^* de E, esto es, abreviadamente, $H : \theta \in E^*$.

La hipótesis que se desea contrastar se denomina hipótesis nula[36], y se escribe simbólicamente como $H_0 : \theta \in E_0$, donde E_0 es un subconjunto de E.

La negación de la hipótesis nula se denomina hipótesis alternativa, o más brevemente, alternativa, y se escribe simbólicamente como $H_1 : \theta \in E_1$, donde E_1 es el complementario de E_0 en E, esto es, $E_1 = E - E_0$.

Llamemos $\aleph$ al conjunto de resultados posibles de una muestra de tamaño que sea representativa de la población.

Un test de contraste de la hipótesis nula $H_0 : \theta \in E_0$, frente a la alternativa $H_1 : \theta \in E_1$, es una partición[37] de X en dos subconjuntos, A y C. A se denomina región de aceptación, y C,

35. Los contrastes de hipótesis utilizan una terminología muy precisa que es propia de esta teoría y diferente de la utilizada en otras ramas de la Estadística. Es necesario que el profesor se sitúe en la misma y para ello puede recurrir a manuales como Ruiz Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. Opus cit. 1995. Cap. 8), Casas (CASAS, J. *Inferencia Estadística para economía y administración de empresas*. Centro de Estudios Ramón. Areces. Madrid, 1996. Págs. 297 a 308) o Peña (PEÑA, D. *Fundamentos de Estadística*. Alianza Editorial. Madrid, 2001. Cap. 10). En Llopis (LLOPIS PÉREZ, J. *La estadística: una orquesta hecha instrumento*. Ariel Ciencia. Barcelona, 1996. Págs. 268 a 348) se encuentra un tratamiento intuitivo del problema, que puede ser útil a quienes se acerquen por primera vez al mismo.

36. Algunos autores (por ejemplo, Peña (PEÑA, D. Opus cit. 2001. Pág. 382), haciendo suyo un comentario clásico), comparan la lógica de los contrastes "*... a la de un juicio penal, donde debe decidirse si el acusado es inocente o culpable. La hipótesis (nula) es que el acusado es inocente y el juicio consiste en aportar evidencia suficiente para rechazar esta hipótesis de inocencia más allá de cualquier duda razonable*"

37. La palabra "*nula*" aplicada a la hipótesis que se desea contrastar tiene su origen en un ejemplo de Fisher, denominado el de la dama y el té con leche, en el que se realiza

región de rechazo o región crítica, debiendo entenderse que si $(x_1,...,x_n)$ es un punto de $\aleph$ que pertenece a A, se acepta la hipótesis, y si pertenece a C se rechaza, y por tanto se acepta la alternativa.

Como puede observarse, hay tantos test para contrastar una hipótesis nula frente a una alternativa como maneras de dividir $\aleph$ en dos partes. Ello no quiere decir que todos los test sean igualmente razonables ni igualmente buenos.

Parece lógico juzgar la bondad de un test en relación con la escasez de errores que produzca. Estudiemos la bondad del test propuesto en el ejemplo número 1 anterior.

Ejemplo 1 (continuación). Podemos ahora formalizar la situación planteada en el ejemplo. Como antes decíamos, la población puede describirse mediante una variable aleatoria de Bernoulli,

$$X = \begin{cases} 1 & p[X=1]=p \\ 0 & p[X=0]=1-p \end{cases}$$

donde p, proporción poblacional de alumnos que han ido al cine, caracteriza la distribución de probabilidades de X. Su campo de variación, a falta de otra información, es $E=[0,1]$.

Con la notación anterior, la hipótesis nula será $H_0 : p \in [0, 0{,}25]$, esto es, $E_0 = [0, 0{,}25]$. La alternativa se escribe como $H_1 : p \in (0{,}25, 1]$ si utilizamos la notación general, es decir, en este ejemplo, $E_1 = (0{,}25, 1]$.

Más arriba hemos construido las regiones de aceptación y de rechazo a que nos conduce el criterio de trasladar al estimador lo que

---

un contraste para discernir la habilidad de dicha señora para reconocer si el té se había añadido a la leche o viceversa. La hipótesis que Fisher desea contrastar es su sospecha, esto es, la nula habilidad de la dama para reconocer la diferencia por el sabor. Desde entonces, esta palabra se ha mantenido para calificar la hipótesis inicial propuesta. Pueden verse más comentarios sobre este hecho en Ruiz Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. Opus cit. 1995. Pág. 340).

las hipótesis afirman para el parámetro. Escritas más abreviadamente en función del estimador de p, son

$$A=\left\{\bar{x}=\frac{\sum x_i}{5}\leq 0{,}25\right\} \text{ y } C=\left\{\bar{x}=\frac{\sum x_i}{5}>0{,}25\right\}$$

Los errores que pueden cometerse dependen, por un lado, de la proporción poblacional, y, por otro, de la proporción muestral; así, las situaciones posibles pueden describirse de la siguiente manera:

	H_0 es cierta ($p \leq 0{,}25$)	H_0 es falsa ($p > 0{,}25$)
Aceptamos H_0 ($\bar{x} \leq 0{,}25$)	acierto	error (tipo II)
Rechazamos H_0 ($\bar{x} > 0{,}25$)	error (tipo I)	acierto

Como puede verse, las casillas de la diagonal principal corresponden a aciertos, siendo, por tanto, casillas "deseables". Las de la diagonal secundaria son errores[38]; el de la izquierda rechaza equivocadamente la hipótesis nula, que es cierta; se denomina error de primer tipo o de primera especie. El de la derecha, acepta equivocadamente una hipótesis falsa; recibe el nombre de error de segundo tipo, o de segunda especie. Como ignoramos el valor del parámetro, no podemos decir en qué columna nos encontramos (pero estamos en una de las dos)

Como antes decíamos, un buen test será aquel que produzca pocos errores. Idealmente, debiéramos elegir una región de aceptación, A, y una crítica, C, tal que, si la hipótesis es cierta, nunca la rechazásemos, esto es, si el valor (desconocido) de p no es superior a 0,25, ninguna de las posibles muestras debiera estar en C. Con ello, nunca cometeríamos un error de tipo I.

38. Obviamente, la clasificación de los errores en tipos depende de la hipótesis que sea escogida como nula. Esta es una decisión de gran importancia, a la que los textos no dedican en general atención. Puede intuirse este hecho en el apartado sobre la asimetría en los contrastes de hipótesis, en este mismo capítulo.

Además, dichas regiones debieran cumplir que, si la hipótesis fuese falsa siempre sería rechazada, o sea, si $p > 0{,}25$, todas las posibles muestras estuvieran en C; el error de tipo II no se cometería.

Conjugar ambas pretensiones es imposible ya que casi cualquier resultado muestral es posible para cualquier valor de p. Debemos, por tanto, rebajarlas, exigiendo sólo que el error de tipo I se cometa para una pequeña "proporción" de muestras, y que otro tanto ocurra con el de tipo II.

Para simplificar, de momento, el problema, supongamos que los 5 alumnos se eligen mediante 5 sorteos, de manera que un alumno podría aparecer varias veces en una muestra (es como escoger 5 bolas con reemplazamiento en una urna)

Supongamos de momento que la hipótesis es cierta, por ejemplo, que la proporción poblacional, p, vale 0,1. Cuando se define lo que es una probabilidad, se insiste en que la probabilidad de un suceso no debe ser muy diferente de la frecuencia (proporción) con que dicho suceso ocurre. Por tanto, la proporción de muestras para las que rechazamos la hipótesis, esto es, para las que $\bar{x} = \frac{x_1 + \cdots + x_5}{5} > 0{,}25$, será la probabilidad de que esto ocurra, esto es, $\Pr[\bar{X} = (X_1 + \cdots + X_5)/5 > 0{,}25]$.

Pero esta probabilidad se puede escribir como[39]

$$\Pr[(X_1 + \cdots + X_5) > 1{,}25] = \Pr[(X_1 + \cdots + X_5) \geq 2] = 1 - \Pr[(X_1 + \cdots + X_5) \leq 1]$$

donde la primera igualdad se deduce del hecho de que $X_1 + ... + X_5$ sólo toma valores enteros, y, como p vale 0,1, $(X_1 + \cdots + X_5)$es una variable aleatoria binomial, $B(5, 0{,}1)$, esto es, mide el número de éxitos en cinco tiradas de Bernoulli, independientes, con probabilidad de éxito 0,1. Si acudimos a las tablas de esta variable,

$$Pr[\bar{X} > 0{,}25] = 1 - 0{,}9185 = 0{,}0815,$$

39. Utilizamos *Pr*, en lugar de *p*, para describir la probabilidad de los sucesos, ya que la letra *p* designa aquí la proporción poblacional.

esto es, si $p = 0{,}1$, aproximadamente para el 8% de las muestras rechazaremos la hipótesis, o sea, concluiremos que $p > 0{,}25$.

Podemos hacer este cálculo para otros valores de la hipótesis. Presentamos los resultados en la siguiente tabla:

p	$Pr[\overline{X} > 0{,}25]$
0,01	0,0006
0,05	0,0140
0,10	0,0815
0,15	0,1095
0,20	0,1808
0,25	0,2617

De la misma forma, podemos calcular la proporción de muestras para las que se acepta la hipótesis siendo p un valor mayor que 0,25, esto es, la envergadura del error de tipo II. Así,

$$\Pr(A) = \Pr[\overline{X} \leq 0{,}25] = \Pr[X_1 + \cdots + X_5 \leq 1{,}25] = \Pr[X_1 + \cdots + X_5 \leq 1]$$

Como $X_1 + \cdots + X_5 \mapsto B(5, p)$, obtenemos la siguiente tabla:

p	$\Pr[\overline{X} \leq 0{,}25]$	p	$\Pr[\overline{X} \leq 0{,}25]$
0,3	0,5282	0,65	0,0540
1/3	0,4609	2/3	0,0453
0,35	0,4284	0,7	0,0308
0,4	0,3370	0,75	0,0156
0,45	0,2562	0,8	0,0067
0,5	0,1875	0,85	0,0022
0,55	0,1312	0,9	0,0005
0,6	0,0870	0,95	0,0000
		0,99	0,0000

Lo que las tablas anteriores muestran ayuda a evaluar la envergadura de los errores que se cometen cuando el criterio adoptado con-

siste en comprobar si la proporción muestral verifica la hipótesis o no la verifica (esto es, cuando se toma como región crítica $C = \{\bar{x} > 0{,}25\}$). Así, si la hipótesis es cierta, la rechazamos para proporciones de muestras que van entre el 0% y el 26,17%. Si es falsa, la proporción de muestras para las que la aceptaremos llega hasta el 52,82%.

Más adelante evaluaremos la importancia de estos errores y el modo de disminuirlos. De momento, y para fijar ideas, generalizaremos este modo de trabajar.

Supongamos que estamos contrastando la hipótesis nula $H_0 : \theta \in E_0$, frente a la alternativa $H_1 : \theta \in E_1$, donde $E = E_0 \cup E_1$ es el conjunto de los valores posibles del parámetro. Sea $\aleph$ el conjunto de los resultados posibles de una muestra representativa de tamaño n. Sea ϕ un test de región crítica C, y de aceptación, A.

Decimos que hemos cometido un error de tipo I cuando rechazamos la hipótesis nula siendo cierta, esto es, cuando $\theta \in E_0$ y $(x_1,...,x_n) \in C$. Asimismo, decimos que hemos cometido un error de tipo II cuando aceptamos la hipótesis nula siendo falsa, es decir, cuando $\theta \in E_1$ y $(x_1,...,x_n) \in A$. Definimos $\alpha(\theta), \theta \in E_0$ como la probabilidad del error de tipo I,

$$\alpha(\theta) = p_\theta(C), \theta \in E_0$$

Definimos $\beta(\theta), \theta \in E_1$ como la probabilidad del error de tipo II,

$$\beta(\theta) = p_\theta(A) = 1 - p_\theta(C), \theta \in E_1.$$

Como puede observarse, $\alpha(\theta)$ y $\beta(\theta)$ dependen del test elegido; de hecho, algunos libros los denotan, si el test se escribe como ϕ, en la forma $\alpha_\phi(\theta)$ y $\beta_\phi(\theta)$ para poner de manifiesto esta dependencia. Nosotros no lo haremos, salvo cuando sea estrictamente necesario.

Nótese, además, que $\alpha(\theta)$ y $\beta(\theta)$ no son comparables, ya que están definidas en distintos (y disjuntos) subconjuntos de E. La función de potencia del test ϕ engloba, en cierto modo, a ambas:

Dado un test, ϕ, de región crítica C, se define su función de potencia, π_ϕ o π, si no hay confusión, como la aplicación $\pi : E \to \Re$, dada por

$$\pi(\theta) = p_\theta(C), \theta \in E.$$

La función de potencia, como hemos dicho, describe, en cierto modo, las probabilidades de ambos tipos de error. Nótese que

$$\pi(\theta)=p_\theta(C)=\begin{cases}\alpha(\theta) & \theta\in E_0\\ 1-p_\theta(A)=1-\beta(\theta) & \theta\in E_1\end{cases}$$

Aunque desde el punto de vista formal la función de potencia parece un concepto de interés, en la práctica no es así; en el resto del trabajo no la volveremos a utilizar. De hecho, no debe confundirse con lo que se denomina *potencia* de un test; este concepto sí que será utilizado. Como veremos, la potencia de un test es una "porción" de la función de potencia.

Dado un test, ϕ, de región crítica C, se define su potencia como

$$1-\beta(\theta)=p_\theta(C),\theta\in E_1$$

Con la jerga que acabamos de introducir, nuestro interés será elegir test para los que $\alpha(\theta)$y $\beta(\theta)$ sean pequeñas, esto es, para los que la función de potencia sea pequeña en E_0 y grande en E_1. No es sencillo conseguir ambos objetivos simultáneamente, ya que, con frecuencia, reducir $\alpha(\theta)=p_\theta(C),\theta\in E_0$, obliga a reducir la región crítica, C, con lo que también se reduce $1-\beta(\theta)=p_\theta(C),\theta\in E_1$, aumentando en consecuencia $\beta(\theta),\theta\in E_1$. Se impone una solución de compromiso; la que más éxito ha tenido hasta la fecha se debe a los estadísticos Neyman y Pearson, y consiste en limitar a un valor razonable, α, la probabilidad de error de tipo I, y buscar, entre los test que cumplen esta limitación, aquel que más disminuye la probabilidad de error de tipo II, esto es, el que maximiza la potencia. Formalicemos este criterio.

Sea α un número del abierto $(0,1)$. Se dice que un test ϕ, de región crítica C, es de nivel de significación α si cumple que

$$\alpha_\phi(\theta)\le\alpha,\ \theta\in E_0.$$

Se dice que ϕ es el test uniformemente más potente al nivel de significación α si maximiza la potencia entre todos los de dicho nivel de significación. Formalmente, debe cumplir

1. $\alpha_{\phi}(\theta) \leq \alpha, \theta \in E_0$
2. **Si ϕ' es un test tal que $\alpha_{\phi'}(\theta) \leq \alpha, \theta \in E_0$, se cumple**

$$\beta_{\phi}(\theta) \leq \beta_{\phi'}(\theta), \quad \theta \in E_1$$

Obsérvese que el nivel de significación es una cota superior de la probabilidad de error de tipo I, y que debe ser fijada por la persona que realiza el test. Un test de nivel de significación, digamos, 0,1, rechazará hipótesis ciertas para no más del 10% de las posibles muestras. En líneas generales, cuanto más pequeño sea α, más pequeña tendrá que ser la región crítica, y mayor será la probabilidad de error de segundo tipo; para una mayor proporción de resultados muestrales se aceptarán hipótesis falsas.

En definitiva, α debe elegirse como el **mayor** valor permisible para la probabilidad de error de tipo I y debe, además, apurarse la desigualdad $\alpha(\theta) \leq \alpha, \theta \in E_0$.

La práctica común, no obstante, consiste en elegir valores muy pequeños de α que conducen a valores muy pequeños de $\alpha(\theta), \theta \in E_0$, y muy grandes de $\beta(\theta), \theta \in E_1$. Se rechazan, por tanto, pocas veces hipótesis ciertas, pero se aceptan frecuentemente hipótesis falsas.

En consecuencia, si uno elige un nivel de significación, α, pequeño, y el resultado muestral le conduce a rechazar la hipótesis, puede estar razonablemente confiado en que la hipótesis será falsa (pocas veces se rechazan hipótesis ciertas). Pero si el resultado muestral le conduce a aceptar la hipótesis, la confianza en que realmente sea cierta será pequeña (muchas veces se aceptan hipótesis falsas). Los estadísticos preferimos hablar de que la hipótesis **no se rechaza**, más bien que de su aceptación.

¿Cómo puede conseguirse disminuir, lo más posible, las probabilidades de ambos tipos de errores? No hay una respuesta general a esta pregunta[40], pero sí se pueden sugerir ciertas pautas.

40. Aunque en este material didáctico las regiones de aceptación y crítica se construyen con criterios intuitivos, existe un teorema, denominado de Neyman y Pearson que propor-

1. Utilícese, para construir la región de aceptación y la crítica, un buen estimador de θ. A ser posible, uno que esté poco disperso en torno al parámetro.
2. Constrúyase la región crítica imponiendo al estimador condiciones similares a las que la alternativa impone al parámetro.
3. Obténgase una muestra del mayor tamaño posible; el principio general es que, a mayor información, menores errores.

Veamos cómo se concretan estas ideas para nuestros ejemplos.

Ejemplo 1 (continuación). Revisemos el ejemplo 1 a la luz de los anteriores resultados y definiciones[41].

En primer lugar, hablemos del estimador. Hemos elegido como estimador de la proporción poblacional la proporción muestral. La calidad de este estimador ya fue ponderada en el estudio de la Inferencia estadística, y no parece mejorable. Asimismo, hemos seguido la segunda pauta a rajatabla, imponiendo al estimador en la región crítica **la misma** condición que la hipótesis alternativa impone al parámetro. Fijemos uno de los niveles de significación "habituales", por ejemplo, α=0,1 (son frecuentes α= 0,1, 0,05 o 0,01) Ello implica que exigimos que, si la hipótesis es cierta (si $p \leq 0{,}25$), no más de un 10% de los resultados muestrales nos conducirán a su rechazo por pertenecer a la región crítica.

Con la ayuda de un instrumento como *DERIVE*, o bien una hoja de cálculo, podemos dibujar las probabilidades de ambos tipos de error para cada valor de p. Como

---

ciona una técnica de obtención de regiones críticas cuyas propiedades estadísticas son, en cierto modo, óptimas en una diversidad de casos e hipótesis. Puede verse demostrado en, por ejemplo, Casas (Opus cit. 1996. Pág. 357) o en Ruiz Maya y Martín Pliego (Opus cit. 1995. Pág. 354). Aunque el resultado se demuestra para hipótesis y alternativa simples, se amplía sin dificultad a situaciones más generales.

41. Este ejemplo propone un contraste de los denominados "*de proporciones*", en los que la hipótesis versa sobre una proporción poblacional y utiliza en la contrastación la correspondiente proporción muestral. Los resultados generales sobre estos contrastes pueden verse en Peña (PEÑA, D. Opus cit. 2001. Pág. 391), Casas (CASAS, J. Opus cit. 1996. Pág. 452), o Ruiz Maya y Martín Pliego (RUIZ MAYA, L. Y MARTÍN PLIEGO, F. Opus cit. 1995. Pág. 516), entre otros.

$$\begin{aligned} \Pr_p(C) &= \Pr\left[\overline{X} > 0{,}25\right] = \Pr\left[\sum X_i > 1{,}25\right] = 1 - \Pr\left[\sum X_i \leq 1{,}25\right] = \\ &= 1 - \left(\Pr\left[\sum X_i = 0\right] + \Pr\left[\sum X_i = 1\right]\right) = \\ &= 1 - \left[(1-p)^5 + 5p(1-p)^4\right], \end{aligned}$$

la probabilidad de error de primer tipo es una parte de dicha función,

$$\alpha(p) = 1 - \left[(1-p)^5 + 5p(1-p)^4\right], \quad p \in [0, 0{,}25]$$

y la potencia del test la parte restante,

$$1 - \beta(p) = 1 - \left[(1-p)^5 + 5p(1-p)^4\right], \quad p \in (0{,}25, 1]$$

Así, la probabilidad de error de segundo tipo resulta

$$\beta(p) = (1-p)^5 + 5p(1-p)^4, \; p \in (0{,}25, 1].$$

Como puede verse en la gráfica que presentamos más adelante, el test no es de nivel de significación 0,1, ya que la probabilidad de error de primer tipo rebasa ese valor (la tabla construida anteriormente indica también este hecho; por ejemplo, si $p = 0{,}20$, uno de los casos en que la hipótesis es cierta, $\alpha(0{,}20) = 0{,}1808$, por lo que para un 18% de resultados muestrales se rechazaría la hipótesis).

Además, como puede verse tanto en la tabla como en la gráfica, la potencia del test no es muy alta. Para $p = 0{,}3$, la probabilidad de error de segundo tipo es 0,5282 (potencia $\pi(0{,}3) = 1 - \beta(0{,}3) = 0{,}4718$), por lo que el 52,82% de los resultados muestrales llevarán a la aceptación de la hipótesis, pese a ser falsa.

Si queremos respetar la restricción que impone el nivel de significación, tendremos que modificar el test (la región crítica). Parece razonable seguir manteniendo su aspecto, esto es, tomar

$$C = \{\overline{x} > r\},$$

ya que, si para una proporción muestral rechazamos $H_0 : p \leq 0{,}25$, también debiéramos rechazar la hipótesis para proporciones muestrales superiores.

El valor adecuado de r es el que impone el nivel de significación, mediante la restricción $\alpha(p) \leq 0{,}1,\ p \in [0{,}0{,}25]$. Como

$$\begin{aligned} \alpha(p) &= \Pr[\overline{X} > r] = \Pr[\sum X_i > 5r] = \\ &= \Pr[B(5,p) > 5r] = 1 - \Pr[B(5,p) \leq 5r] \end{aligned}$$

la restricción se concreta en[42] $\Pr[B(5,p) \leq 5r] \geq 0{,}9$.

Como ilustran las tablas de la distribución binomial, basta con imponer esa condición para el "peor" valor de la hipótesis, esto es, para el punto de la hipótesis que más se parece a la alternativa, que es $p = 0{,}25$.

Resulta, pues,

$$\Pr[B(5,0{,}25) \leq 5r] \geq 0{,}9,$$

de donde $5r = 3$, ya que, en las tablas de dicha distribución

$$\Pr[B(5,0{,}25) \leq 3] = 0{,}9844 \geq 0{,}9.$$

En este caso, por tanto,

$$\alpha(0{,}25) = 1 - 0{,}9844 = 0{,}0156.$$

Como se ve, estamos dispuestos a rechazar hipótesis ciertas un 10% de las veces y, para 0,25, sólo las rechazamos un 1,5%. Esto es inevitable cuando la distribución de probabilidad del estimador es discreta. En cualquier caso, recordemos que la desigualdad $\alpha(p) \leq 0{,}1,\ p \leq 0{,}25$, debe intentarse apurar todo lo posible, llegando, incluso, a la igualdad.

Al haber impuesto la condición para el "peor" valor de la hipótesis, es razonable que se cumpla con más holgura para los demás. En efecto, la región crítica obtenida es

$$C' = \left\{ \sum x_i > 3 \right\} = \left\{ \overline{x} > \frac{3}{5} \right\},$$

42. Escribimos Pr[B(5,p) < 5r] para referirnos de forma abreviada a la probabilidad de que una variable, Z, binomial B(5,p), cumpla Z < 5r.

por lo que la condición $\alpha(p) \leq 0{,}1, p \leq 0{,}25$ obliga a calcular

$$\alpha(p) = \Pr(C') = \Pr[B(5, p) > 3] = 1 - \Pr[B(5, p) \leq 3].$$

La siguiente tabla muestra los valores obtenidos para los valores de la hipótesis que aparecen en las tablas de la distribución binomial:

p	0,01	0,05	0,1	0,15	0,20	0,25
$\alpha(p)$	0,0000	0,0000	0,0005	0,0022	0,0067	0,0156

Puede compararse la región C' con C. La idea básica sigue siendo la misma (se rechaza la hipótesis si la proporción muestral es grande), pero la reducción de la probabilidad de error de tipo I impone la reducción de la región crítica. Ahora, para rechazar la hipótesis, es necesario que la proporción muestral sea muy grande; en concreto, superior al 60% (antes esta cifra era del 25%).

¿Qué ha ocurrido con la probabilidad de errores de tipo II y con la potencia del test? La siguiente tabla muestra sus valores:

p	$\beta(p)$	$1-\beta(p)$	p	$\beta(p)$	$1-\beta(p)$
0,3	0,9692	0,0308	0,65	0,5716	0,4284
1/3	0,9547	0,0453	2/3	0,5391	0,4609
0,35	0,9460	0,0540	0,7	0,4718	0,5282
0,4	0,9130	0,0870	0,75	0,3672	0,6328
0,45	0,8688	0,1312	0,8	0,2627	0,7373
0,5	0,8125	0,1875	0,85	0,1648	0,8352
0,55	0,7438	0,2562	0,9	0,0815	0,9185
0,6	0,6630	0,3370	0,95	0,0226	0,9774
			0,99	0,0010	0,9990

En definitiva, no se garantizan potencias altas, aunque varían con el valor de p; si, por ejemplo, p vale 0,3, se aceptaría que la hipótesis es cierta (no lo es, obviamente) para un 96,3% de los resultados muestrales. No todas las situaciones que describe la tabla anterior son, claro está, tan catastróficas.

La siguiente gráfica muestra los efectos que sobre la función de potencia ha tenido el cambio del test ϕ (de región crítica C) al ϕ' (de región crítica C').

En la encuesta se obtuvieron tres respuestas afirmativas (el resultado fue, recordemos, SNSSN) Este resultado puede escribirse como

$$(x_1,...,x_5)=(1,0,1,1,0),$$

por lo que podemos calcular la proporción muestral

$$\bar{x}=\frac{x_1+...+x_5}{5}=\frac{3}{5}=0{,}6\,.$$

Usando el test ϕ', de región crítica $C'=\{\bar{x}>0{,}6\}$, concluimos que no puede rechazarse la hipótesis, ya que $(1,0,1,1,0)\notin C'$.

¿Será cierta o falsa la hipótesis?, esto es, ¿habrán ido al cine no más del 25 %? No es posible responder a esta pregunta sin hacer un censo, o sea, sin preguntar a todo el Instituto. El test que hemos realizado nos dice que se aceptan hipótesis falsas en una alta proporción, que puede llegar hasta el 96,9% de los resultados muestrales si p vale 0,3, por ejemplo. Es decir, cabe esperar que, si $p=0{,}3$, aceptaríamos la hipótesis en 97 de cada 100 encuestas realizadas. Recuérdese, en este sentido, que la baja potencia a que conduce tomar un nivel de significación, α, bajo, no permite aceptar hipótesis sino, más bien, no rechazarlas.

Como dijimos anteriormente, los errores pueden disminuirse aumentando la información; si, por ejemplo, encuestásemos a 15 alumnos del Instituto, ahora el conjunto de resultados muestrales sería

$$\aleph=\left\{\left(x_1,...,x_{15}\right),x_i=0,1,i=1,...,15\right\}$$

y la región crítica,

$$C''=\left\{\left(x_1,...,x_{15}\right)\in\aleph,\ \text{tales que}\ \bar{x}>r\right\}$$

o, abreviadamente,

$$C''=\{\bar{x}>r\}.$$

Obtengamos el valor actual de r. Puesto que vamos a comparar esta región con C', impongamos un nivel de significación del orden del error de tipo I anterior, que era $\alpha_{\phi'}(0{,}25)=0{,}0156$. Ahora, llamando ϕ'' al test cuya región crítica es C'',

$$\begin{aligned}\alpha_{\phi''}(0{,}25) &= \Pr[\bar{X}>r]=\Pr[\textstyle\sum X_i>15r]= \\ &= \Pr[B(15,0{,}25)>15r]=1-\Pr[B(15,0{,}25)\leq 15r],\end{aligned}$$

por lo que tomaremos, a la vista de las tablas, $15r=7$, ya que entonces

$$\alpha_{\phi''}(0{,}25)=1-\Pr[B(15,0{,}25)\leq 7]=1-0{,}9827=0{,}0173$$

muy similar al anterior.

La región crítica sería ahora

$$C''=\{\textstyle\sum x_i>7\}=\left\{\bar{x}>\dfrac{7}{15}\right\}.$$

Obsérvense las diferencias con el resultado obtenido para 5 alumnos encuestados. Antes, con ϕ', necesitábamos obtener una proporción muestral igual o superior al 60 % para rechazar la hipótesis. El aumento de información se traduce en que ahora rechazamos $H_0: p\leq 0{,}25$ si obtenemos una proporción muestral superior a sólo el 46,7 % (7/15=0,467), y ello con un valor similar del nivel de significación **efectivo** (antes era del 1,56 %, y ahora del 1,73 %).

Pero, además, ha aumentado la potencia del test. Así, si calculamos $\beta_{\phi''}(0{,}3)$,

$$\beta_{\phi''}(0{,}3)=\Pr[\textstyle\sum X_i\leq 7]=\Pr[B(15,0{,}3)\leq 7]=0{,}95.$$

Con el test ϕ', si p vale 0,3, aceptábamos la hipótesis para el 96,9 % de los resultados muestrales. Con ϕ'', sólo ocurre para el 95 %. Las diferencias son más notorias para valores superiores de p, como muestra la gráfica anterior.

Si el número de encuestados es superior a 20, habrá que utilizar la aproximación normal a la distribución binomial. El ejemplo se desarrollaría de forma similar, con la "ventaja" de que ahora podríamos apurar el nivel de significación, esto es,

$$\text{máx}_{p\leq 0,25}\alpha(p)=\alpha(0,25)=\alpha$$

Completemos ahora el ejemplo 2.

Ejemplo 2 (continuación). En este ejemplo se deseaba contrastar la hipótesis de que la altura media de los alumnos de Bachillerato rebasa 1,70 metros. Desarrollaremos el contraste, utilizando la misma técnica que empleamos para el ejemplo 1.

En este caso, se acepta que la altura de los alumnos, X, sigue una distribución $N(\mu,\sigma)$. La hipótesis nula es $H_0:\mu\geq 1{,}70$, y la alternativa, $H_1:\mu<1{,}70$. Con el criterio adoptado anteriormente, la región crítica sería

$$C=\{\ (x_1,x_2,x_3)\ \text{tales que}\ \bar{x}<1{,}70\}$$

o, abreviadamente, $C=\{\bar{x}<1{,}70\}$ ya que, recordemos, rechazábamos la hipótesis si la estimación de la altura media poblacional, que es la altura media del resultado muestral, $\bar{x}=\dfrac{x_1+x_2+x_3}{3}$, cumplía la condición de la alternativa. La región de aceptación sería $C=\{\bar{x}\geq 1{,}70\}$.

Obtengamos las probabilidades de error de ambos tipos y la potencia del test.

La probabilidad de error de primer tipo es la función

$$\alpha(\mu)=p_\mu(C),\ \mu\geq 1{,}70$$

donde, con $p_\mu(C)$ queremos indicar la probabilidad de la región crítica cuando la media poblacional es igual a μ.

Ahora bien, $p_\mu(C)=p_\mu\left[\overline{X}<1{,}70\right]$. Recordemos que, cuando estudiamos las técnicas de estimación de medias de poblaciones normales, indicamos, sin demostrarlo, que la media muestral se distribuía también como una normal; en concreto,

$$\overline{X}\mapsto N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$$

en nuestro caso $\overline{X}\mapsto N\left(\mu,\frac{\sigma}{\sqrt{3}}\right)$. Por tanto, dicha probabilidad puede calcularse tipificando la media muestral, a saber,

$$\alpha(\mu)=p_\mu\left[\frac{\overline{X}-\mu}{\sigma/\sqrt{3}}<\frac{170-\mu}{\sigma/\sqrt{3}}\right]=p_\mu\left[N(0,1)<\frac{170-\mu}{\sigma/\sqrt{3}}\right]$$

ya que $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{3}}$ tiene una distribución $N(0,1)$.

Es posible obtener $\alpha(\mu)$ para cualquier valor de μ. No obstante, la técnica de construcción de test que estamos desarrollando, sólo evalúa la máxima probabilidad de error de primer tipo. Calculémosla para nuestro test.

Según crece μ, decrece $(1{,}70-\mu)/(\sigma/\sqrt{3})$, luego decrece también $\alpha(\mu)$. Por tanto, la probabilidad de error de primer tipo, $\alpha(\mu), \mu\geq 1{,}70$, será menor o igual que $\alpha(1{,}70)$.

En definitiva, para la hipótesis nula (para $\mu\geq 1{,}70$)

$$\alpha(\mu) \leq \alpha(1{,}70)=p_{1{,}70}\left[N(0,1)<\frac{1{,}70-\mu}{\sigma/\sqrt{3}}\right]=$$

$$= p\left[N(0,1)<\frac{1{,}70-1{,}70}{\sigma/\sqrt{3}}\right]=p\left[N(0,1)<0\right]=0{,}5.$$

El test que tiene región crítica $C=\{\overline{x}<1{,}70\}$ tiene un nivel de significación 0,5; rechaza, por tanto, hipótesis ciertas para no más del 50 % de los resultados muestrales. De hecho, si la altura media fuese

exactamente 1,70 metros, rechazaría la hipótesis para la mitad de los resultados muestrales, aproximadamente.

No obstante, la práctica sugiere tomar niveles de significación sensiblemente más pequeños que 0,5. Tomemos, como en el ejemplo 1, un nivel de significación $\alpha=0{,}1$, y llamemos, para abreviar, ϕ al test anterior, que tenía como región crítica C. Parece razonable seguir rechazando la hipótesis para valores pequeños de la media muestral, ya que la alternativa especifica valores pequeños de la media poblacional ($\mu<1{,}70$). En definitiva, tomaremos una región crítica de la forma

$$C'=\{\bar{x}<k\}$$

Si ahora escribimos la probabilidad del error de primer tipo, tipificando la media muestral como antes, obtendremos

$$\begin{aligned}\alpha(\mu) &= p_\mu[\bar{X}<k]=p_\mu\left[\frac{\bar{X}-\mu}{\sigma}\sqrt{3}<\frac{k-\mu}{\sigma}\sqrt{3}\right]= \\ &= p\left[N(0,1)<\frac{k-\mu}{\sigma}\sqrt{3}\right]\end{aligned}$$

También como antes, al crecer μ decrece $\alpha(\mu)$, luego la condición $\alpha(\mu)\leq 0{,}1$, $\mu\geq 1{,}70$ se satisface (y la desigualdad se apura) si tomamos k como el valor tal que

$$\alpha(1{,}70)=p\left[N(0,1)<\frac{k-1{,}70}{0{,}15}\sqrt{3}\right]=0{,}1\,.$$

Si σ es conocida, esta ecuación permitirá obtener el valor de k. Por ejemplo, si vale 15 centímetros ($\sigma=0{,}15$), resulta

$$p\left[N(0,1)<\frac{k-1{,}70}{0{,}15}\sqrt{3}\right]=0{,}1\,,$$

con lo que, a través de las tablas de la $N(0,1)$ obtenemos

$$\frac{k-1{,}70}{0{,}15}\sqrt{3}=-1{,}28.$$

En consecuencia,

$$k = 1{,}70 - \frac{0{,}15 \cdot 1{,}28}{\sqrt{3}} = 1{,}59 \text{ metros.}$$

La región crítica resulta

$$C' = \{\bar{x} < 1{,}59\}.$$

Puesto que la altura media de los tres encuestados ha sido de 1,75 metros, no rechazaríamos la hipótesis, ya que $(1{,}75,\ 1{,}80,\ 1{,}70) \notin C'$.

En la mayor parte de las aplicaciones, la desviación típica poblacional, σ, es desconocida. Como vimos en el capítulo anterior, dedicado a los intervalos de confianza, este problema se resuelve sustituyendo σ por una estimación suya, la cuasi-desviación típica muestral, s_c, raíz cuadrada positiva de la cuasi-varianza muestral. La consecuencia de esta sustitución es que, si la muestra es pequeña (digamos, no superior a 31 encuestados), el valor de $\frac{k - 1{,}70}{s_c}\sqrt{3}$ se debe buscar en las tablas de otra variable, la t de Student con $n-1$ grados de libertad, donde n es el tamaño de la muestra.

En nuestro caso, hay 3 - 1 = 2 grados de libertad, por lo que

$$\frac{k - 1{,}70}{s_c}\sqrt{3} = -1{,}89$$

y, por tanto,

$$k = 1{,}70 - \frac{1{,}89 \cdot s_c}{\sqrt{3}}.$$

Calculemos s_c.

$$s_c^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2}\left[(1{,}75 - 1{,}75)^2 + (1{,}80 - 1{,}75)^2 + (1{,}70 - 1{,}75)^2\right] =$$
$$= \frac{1}{2} \cdot 0{,}005 = 0{,}0025,$$

de donde $s_c = \sqrt{0{,}0025} = 0{,}05$.

En definitiva,

$$k = 1{,}70 - \frac{1{,}89 \cdot 0{,}05}{\sqrt{3}} = 1{,}65.$$

La región crítica sería

$$C' = \{\bar{x} < 1{,}65\}.$$

Como $(1{,}75, 1{,}80, 1{,}70) \notin C'$, la hipótesis no se rechaza.

1.2. Muestreo sin reemplazamiento en los contrastes

Hasta ahora, hemos supuesto que en la realización de la encuesta había reemplazamiento, esto es, que cada encuestado participaba en los siguientes sorteos. Sin embargo, ninguna encuesta se realiza de esta forma; un encuestado no puede serlo posteriormente, ya que se negaría a responder (o traduciría su enfado en respuestas conscientemente erróneas). Desde el punto de vista formal, la imposibilidad de encuestar a un individuo varias veces causa muchos problemas.

Volvamos, por concretar el comentario, sobre el estudio de la proporción de alumnos del Instituto que han ido al cine la pasada semana. Supongamos que en el Centro hay N_1 alumnos que han ido, y N_2 que no han ido (el número total de alumnos es $N = N_1 + N_2$). Llamemos X_1 al resultado correspondiente al primer encuestado, que vale la unidad si ha ido al cine y cero si no ha ido. La distribución de probabilidad de X_1 es sencilla de obtener,

$$p[X_1 = 1] = p, \text{ y } p[X_1 = 0] = 1 - p,$$

donde p es la proporción poblacional, $p = \frac{N_1}{N}$.

Llamemos X_2 al resultado correspondiente al segundo encuestado. Si hay reemplazamiento, la proporción poblacional sigue siendo la misma tras la primera encuesta, con lo que la distribución de probabilidades de X_2 es la misma que la de X_1,

$$p[X_2 = 1] = p, \text{ y } p[X_2 = 0] = 1 - p,$$

donde, nuevamente, p es la proporción poblacional, $p = \frac{N_1}{N}$. Además, las probabilidades para X_2 no dependen del resultado obtenido para X_1, esto es, X_1 y X_2 son independientes.

Si no hay reemplazamiento, las probabilidades para X_2 dependen del resultado obtenido para X_1. Si $X_1 = 1$ (el primer encuestado ha ido al cine), las proporciones cambian entre la primera y la segunda encuesta. En concreto,

$$p[X_2 = 1 | X_1 = 1] = \frac{N_1 - 1}{N - 1}, \text{ y } p[X_2 = 0 | X_1 = 1] = \frac{N_2}{N - 1}.$$

donde las probabilidades anteriores son probabilidades condicionadas por el resultado de la primera encuesta. Si $X_1 = 0$,

$$p[X_2 = 1 | X_1 = 0] = \frac{N_1}{N - 1}, \text{ y } p[X_2 = 0 | X_1 = 1] = \frac{N_2 - 1}{N - 1}.$$

Por tanto, las variables X_1 y X_2 no son independientes.

No obstante, las probabilidades para X_1 y X_2 son, curiosamente, las mismas. En efecto,

$$p[X_2 = 1] = p[X_2 = 1 | X_1 = 1] \cdot p[X_1 = 1] + p[X_2 = 1 | X_1 = 0] \cdot p[X_1 = 0] =$$
$$= \frac{N_1 - 1}{N - 1} \cdot \frac{N_1}{N} + \frac{N_1}{N - 1} \cdot \frac{N_2}{N} = \frac{(N_1 - 1)N_1 + N_1 N_2}{N(N - 1)} = \frac{N_1(N - 1)}{N(N - 1)} = \frac{N_1}{N} = p$$

y, por tanto, $p[X_2 = 0] = 1 - p[X_2 = 1] = 1 - \frac{N_1}{N} = \frac{N_2}{N}$.

En definitiva, si no hay reemplazamiento, $X_1,...,X_n$ no serían variables independientes.

Recuérdese que la anterior construcción de un test para proporciones descansaba en la idea de que $Z=\sum_{i=1}^{n} X_i$ tenía una distribución binomial. Pues bien, ese hecho no se garantiza si los sumandos no son independientes.

En la práctica, no obstante, las poblaciones son suficientemente numerosas como para que la eliminación de un individuo (e incluso la de unos pocos) no modifique prácticamente las proporciones. Si N es grande, en comparación con k, $\frac{N_1-k}{N-k}$ es prácticamente igual a $\frac{N_1}{N}$. Las consecuencias prácticas son que, aproximadamente, todos los resultados que se prueban con reemplazamiento son válidos también cuando dicho reemplazamiento no existe. Se suele considerar que, si el cociente $\frac{n}{N}$, denominado fracción de muestreo, es menor que 0,1, los comentarios anteriores son válidos, y puede trabajarse como si existiera reemplazamiento.

La situación es similar cuando lo que se estudia es la media, μ, de una población normal, $X \mapsto N(\mu,\sigma)$. Se demuestra que la media muestral, $\overline{X}$, sigue una distribución $N\left(\mu,\sigma/\sqrt{n}\right)$, cuando las variables $X_1,...,X_n$, que miden los valores correspondientes a los n encuestados, son todas ellas $N(\mu,\sigma)$ e independientes. Si la población es pequeña, la independencia puede ser inaceptable. Si X es la altura de los individuos de una población, la altura media puede disminuir tras eliminar a los primeros encuestados, si éstos son muy altos. Análogamente a la situación anterior, los resultados son aproximadamente ciertos si la fracción de muestreo es pequeña.

1.3. Asimetría en los contrastes de hipótesis

Los papeles que juegan la hipótesis y la alternativa en los contrastes no son simétricos; si se cambia la hipótesis por la alterna-

tiva, la región de aceptación no pasa a crítica y recíprocamente. Baste para entender esta cuestión con observar que la técnica que estamos desarrollando limita sólo una de las probabilidades de error (la de tipo I), y no es tan restrictiva en lo referente a la de tipo II.

Ilustremos el comentario a través del ejemplo 1. Para contrastar la hipótesis nula $H_0 : p \leq 0{,}25$ frente a la alternativa $H_1 : p > 0{,}25$ hemos obtenido, para una muestra de tamaño 5, la región crítica $C' = \left\{ \sum x_i > 3 \right\}$ (y, por tanto, una región de aceptación $A' = \left\{ \sum x_i \leq 3 \right\}$), al nivel de significación $\alpha = 0{,}1$.

Planteémonos ahora contrastar la hipótesis $H_0 : p > 0{,}25$ frente a la alternativa $H_1 : p \leq 0{,}25$. Digamos, en primer lugar, que cuando el parámetro varía en un continuo, y no toma valores aislados, el o los valores críticos que separan hipótesis nula y alternativa, se suelen asignar a la hipótesis por conveniencias de cálculo. Se contrastaría entonces la hipótesis nula $H_0 : p \geq 0{,}25$ frente a la alternativa $H_1 : p < 0{,}25$. La región crítica sería entonces de la forma $C^* = \left\{ \sum x_i < a \right\}$ y, si el nivel de significación vuelve a ser 0,1, la condición $\Pr\left[C^*\right] \leq 0{,}1$ para $p \geq 0{,}25$ implica

$$\Pr\left[B(5, 0{,}25) < a\right] \leq 0{,}1$$

lo que conduce a tomar $a = 0$, esto es, $C^* = \left\{ \sum x_i < 0 \right\}$, que es una región crítica vacía. Nunca se rechaza, pues, la hipótesis. Como puede verse, A' no coincide con C^*.

1.4. Contrastes de dos caras

Los contrastes que hemos estudiado hasta ahora son de los denominados de *una cara*, en los que la hipótesis y la alternativa son un intervalo (o una semirrecta). Se demuestra que los que hemos construido para la proporción poblacional, o para la media normal, son uniformemente más potentes, en el sentido de la definición anterior (maximizan la potencia para un nivel de significación dado). Muchas veces se contrastan hipótesis del tipo $p = p_0$, o bien $\mu = \mu_0$.

Las alternativas serán, respectivamente, $p \neq p_0$ o bien $\mu \neq \mu_0$. Estos contrastes se denominan de *dos caras*, porque la alternativa tiene ese aspecto ($p < p_0$ ó $p > p_0$, por ejemplo, en el primer caso). En estos casos, puede construirse un test con la misma técnica que la empleada en los ejemplos anteriores, si bien no es de máxima potencia para cada nivel de significación.

Ejemplo 3. Deseando convertirme en tahur, tomo una vieja moneda que, debido al desgaste, tal vez esté cargada. Para ver si la moneda es aprovechable con este fin, intento contrastar si la moneda es perfecta, y para ello la lanzo 40 veces, obteniendo 12 caras. Realizar el contraste adecuado, a un nivel de significación $\alpha = 0{,}1$.

Solución. La hipótesis a contrastar es $H_0 : p = 0{,}5$, y la alternativa, $H_1 : p \neq 0{,}5$. Parece razonable rechazar la hipótesis si la proporción de caras obtenida es muy grande o muy pequeña. En consecuencia, la región crítica será

$$C = \{ \hat{p} < a \text{ ó } \hat{p} > b \},$$

donde $a < b$ son números adecuados. El nivel de significación exigido impone que

$$\Pr_{p=0,5}[C] = \Pr_{p=0,5}[\hat{p} < a] + \Pr_{p=0,5}[\hat{p} > b] \leq 0{,}1.$$

Aunque hay múltiples elecciones de a y b para las que esa condición se cumple, suele dividirse por la mitad la probabilidad de error. Se determinan entonces a y b mediante las condiciones

$$\Pr_{p=0,5}[\hat{p} < a] \leq \frac{0{,}1}{2} = 0{,}05$$

y

$$\Pr_{p=0,5}[\hat{p} > b] \leq \frac{0{,}1}{2} = 0{,}05$$

Como la muestra es de tamaño $n = 40$, las ecuaciones no pueden resolverse con las tablas de la distribución binomial, como

hemos hecho antes para $n=5$. Se acude entonces a la aproximación normal, que indica que, si la proporción poblacional es p

$$\sum_{i=1}^{40} X_i \approx N\left(40\cdot p, \sqrt{40\cdot p\cdot(1-p)}\right),$$

y

$$\hat{p}=\frac{\sum_{i=1}^{40} X_i}{40} \approx N\left(p, \sqrt{p\cdot(1-p)/40}\right).$$

La primera ecuación es, pues,

$$\Pr_{p=0,5}[\hat{p}<a] = \Pr\left[\frac{\hat{p}-0,5}{\sqrt{0,5\cdot 0,5/40}} < \frac{a-0,5}{\sqrt{0,5\cdot 0,5/40}}\right] =$$

$$= \Pr\left[N(0,1) < \frac{a-0,5}{\sqrt{0.5\cdot 0.5/40}}\right] \leq 0,05$$

ecuación que, resuelta para la igualdad, proporciona el valor

$$\frac{a-0,5}{\sqrt{0,5\cdot 0,5/40}} = -1.64,$$

de donde

$$a = 0,5 - \frac{1,64\cdot 0,5}{\sqrt{40}} = 0,5 - 0,13 = 0,37.$$

Análogamente, la ecuación para b es

$$\Pr_{p=0,5}[\hat{p}>b] = \Pr\left[N(0,1) > \frac{b-0,5}{\sqrt{0,5\cdot 0,5/40}}\right] \leq 0,05$$

cuya solución es

$$\frac{b-0,5}{\sqrt{0,5\cdot 0,5/40}} = 1.64,$$

y, por tanto,

$$b = 0{,}5 + \frac{1{,}64 \cdot 0{,}5}{\sqrt{40}} = 0{,}5 + 0{,}13 = 0{,}63.$$

En definitiva, la región crítica es

$$C = \{\, \hat{p} < 0{,}37 \text{ o } \hat{p} > 0{,}63 \}$$

(habría sido más sencillo escribir la de aceptación, $A = \{\, 0{,}37 \leq \hat{p} \leq 0{,}63 \}$).

La proporción muestral obtenida tras los 40 lanzamientos ha sido $\hat{p} = \frac{12}{40} = 0{,}3$, luego la hipótesis de que la moneda está equilibrada se rechaza.

Obviamente, en este ejemplo, no se plantea el problema de que si la muestra lo es con reemplazamiento o sin reemplazamiento. Ello sólo ocurre cuando se encuesta a individuos de una población, aunque sea una población de objetos.

2. UNA PROPUESTA DE UNIDAD DIDÁCTICA

Se quiere estudiar la proporción de alumnos de 2º G de Bachillerato que han repetido al menos un curso a lo largo de su vida escolar. Nos planteamos si esa proporción será de, al menos, el 50%, es decir, que al menos uno de cada dos alumnos ha "perdido" un año.

Previamente se puede haber realizado un censo, dentro de este estudio, o en relación con otro diferente. El profesor conoce la respuesta censal, que describimos en la siguiente tabla:

número de cursos repetidos	*número de alumnos*
0	**10**
1	**11**
2	**8**
3	**1**

Conviene que el alumno sepa que la respuesta existe, es decir, que tratamos de conocer un dato que ignoramos, pero que existe. Es preferible, con todo, que el alumno no conozca la respuesta, para que no oriente intencionadamente su trabajo, ni haga juicios de valor sin fundamento. Nosotros sabemos que esa proporción es $p=\frac{11+8+1}{30}=\frac{2}{3}$**, o sea, el 66,67%.**

La afirmación cuya veracidad queremos estudiar se denomina **hipótesis**, o a veces, hipótesis nula. En nuestro caso, llamando p a la proporción en estudio, la hipótesis puede escribirse como $H_0 : p \geq 0{,}5$. La negación de la hipótesis se denomina **alternativa**, y será que dicha proporción sea menor del 50%, esto es, $H_1 : p < 0{,}5$.

Descartemos la opción censal, y hagamos una encuesta. En general, cuando se encuesta a individuos de la población, cada individuo no puede ser interrogado más de una vez, ya que en ese caso se producirán muchas faltas de respuesta. No obstante, facilita las cosas, como luego veremos, que permitamos que cada individuo pueda ser encuestado tantas veces cuantas sea elegido. De momento, así lo haremos.

Cuando hemos estudiado las técnicas de estimación, hemos insistido en que la muestra elegida para la encuesta tiene que ser representativa de la población. Para ello, de una bolsa que contenga 30 bolas (o 30 papeletas) con numeración correlativa, se escoge una de ellas. Tras anotar su número se devuelve a la bolsa, se agita ésta, y se escoge nuevamente una bola.

Realizamos 5 veces la operación (el tamaño de la muestra, por tanto, será $n=5$). Los alumnos elegidos son los números 8, 25, 10, 23 y 15. Se les pregunta si han repetido al menos un curso; las respuestas son 2 afirmativas y 3 negativas (al haber escrito las respuestas en un papel, se ignora el orden de las cinco respuestas. ¿Estamos perdiendo información?).

En el estudio de las técnicas de estimación, vimos que una estimación razonable de la proporción poblacional, p, era la proporción muestral,

$$\hat{p} = \frac{\text{número de respuestas afirmativas}}{\text{número de individuos encuestados}}$$

Para nuestra muestra, la estimación de p será entonces $\hat{p} = \frac{2}{5} = 0{,}4$.

Intuitivamente, parece que debiera rechazarse la hipótesis, ya que la proporción muestral no la verifica.

La afirmación anterior supone el establecimiento de un criterio para clasificar los resultados muestrales: si, en una encuesta, la proporción muestral es inferior a 0,5, rechazamos la hipótesis, y si es superior o igual a ese valor, la aceptamos. Esta clasificación de los resultados se suele describir agrupando en un conjunto aquellos que conducen a la misma conclusión. Se llama **región de aceptación** al conjunto de los resultados muestrales para los que se acepta la hipótesis. Esta región se denomina con la letra A. En nuestro caso,

$$A = \{\hat{p} \geq 0{,}5\}$$

El conjunto de resultados muestrales para los que se rechaza la hipótesis se denomina región de rechazo o, más frecuentemente, **región crítica**. Se escribe con la letra C. Con nuestro criterio,

$$C = \{\hat{p} < 0{,}5\}$$

Aunque el establecimiento de un criterio para decidir entre hipótesis y alternativa parece tranquilizador, el problema básico no está resuelto. ¿Habremos acertado? ¿será realmente p menor que el 50%? Esta pregunta sólo tiene una respuesta censal; podríamos responderla únicamente si conocemos el valor de p. Pero si no podemos hacer un censo, la pregunta sólo puede responderse en términos relativos, esto es, en términos de riesgo. Y este riesgo viene dado por la mayor o menor frecuencia con que nos equivocamos.

Tomemos, por ejemplo, otra muestra. Los alumnos elegidos son los números 26, 22, 2, 22 y 6, y cuatro de ellos dan una respuesta

afirmativa. Ahora la proporción muestral sería $\hat{p}=\frac{4}{5}=0{,}8$ y, con el criterio anterior, aceptaríamos la hipótesis, ya que $\hat{p}=0{,}8\geq 0{,}5$.

Como puede verse, el criterio dista de ser tranquilizador. Dependiendo del azar, aceptamos o rechazamos la hipótesis. Por ejemplo, nosotros hemos tomado diez muestras más (once, incluyendo la primera), y hemos obtenido las proporciones muestrales 0,8, 1, 0,8, 0,6, 0,6, 1, 0,8, 0,4, 0,8 y 0,4, por lo que, en 2 encuestas aceptaríamos la hipótesis, y en 8 la rechazaríamos.

¿Cómo evaluar el riesgo de rechazar la hipótesis cuando ésta sea cierta? Supongamos por un momento que la proporción poblacional fuese igual a $\frac{2}{3}$, esto es, que 20 de los 30 alumnos hubieran repetido alguna vez. El número de muestras que pueden tomarse es muy elevado ($30^5=24.300.000$), por lo que no es posible inspeccionar todas ellas. Pero la proporción de muestras para las que rechazamos la hipótesis puede evaluarse mediante la probabilidad de que eso suceda (antes de hacer la encuesta, por supuesto). Si llamamos Z al número de encuestados que han repetido algún curso, la proporción muestral es $\hat{p}=Z/5$. Por tanto, la probabilidad de rechazar la hipótesis cuando $p=\frac{2}{3}$ vale

$$\Pr_{p=\frac{2}{3}}[\hat{p}<0{,}5]=\Pr_{p=\frac{2}{3}}\left[\frac{Z}{5}<0{,}5\right]=\Pr_{p=\frac{2}{3}}[Z<2{,}5]=\Pr_{p=\frac{2}{3}}[Z\leq 2]$$

Pero, si $p=\frac{2}{3}$, $Z\mapsto B\left(5,\frac{2}{3}\right)$, luego

$$\Pr_{p=\frac{2}{3}}[\hat{p}<0{,}5]=\Pr\left[B\left(5,\frac{2}{3}\right)\leq 2\right]=1-0{,}7901=0{,}2099$$

Luego, si $p=\frac{2}{3}$, con nuestro criterio (con la región crítica C) tendríamos una probabilidad de 0,2099 de rechazar la hipótesis, esto es, la rechazaríamos para el 20,99% de las muestras.

Puede calcularse esta probabilidad para otros valores de la hipótesis. Algunos resultados los presentamos en la siguiente tabla:

p	0,5	0,6	2/3	0,7	0,8	0,9
$\Pr[\hat{p}<0{,}5]$	0,5	0,3174	0,2099	0,1631	0,0579	0,0086

En concreto, si p vale 0,5 (la hipótesis es cierta, por tanto), la rechazaremos para la mitad de las muestras, mientras que si $p=0{,}9$ (la hipótesis es también cierta), el rechazo se produciría sólo para el 0,86% de las muestras.

En la práctica habitual, riesgos como alguno de los anteriores resultan excesivos, y suelen fijarse valores inferiores para la probabilidad, por ejemplo, 0,1, 0,05 o incluso 0,01, lo que hace inviable nuestro criterio. Esta cantidad recibe el nombre de **nivel de significación del test**, y se suele escribir con la letra griega α.

La idea general, con todo, sigue siendo válida. En efecto, parece razonable mantener que se rechace la hipótesis para valores pequeños de $\hat{p}$.

Esto es, si se rechaza para $\hat{p}=0{,}4$ que $p \geq 0{,}5$, a la misma conclusión debe llegarse para $\hat{p}=0{,}3$, por ejemplo. Por tanto, rechazaremos la hipótesis si $\hat{p} < r$, donde r es un cierto número real. Dicho de otra forma, tomaremos como región crítica

$$C = \{\hat{p} < r\}$$

El número r se determina con la condición de que la probabilidad de rechazar la hipótesis, si es cierta, no rebase el nivel de significación. Veamos cómo hacerlo.

Tomemos un nivel de significación de, por ejemplo, $\alpha=0{,}05$. Llamemos $\Pr_{p_0}[B]$ a la probabilidad de un suceso, B, cuando la proporción poblacional es p_0. Entonces, trabajando de forma similar a la anterior,

$$\Pr_{p_0}[\hat{p} < r] = \Pr_{p_0}\left[\frac{Z}{5} < r\right] = \Pr_{p_0}[Z < 5 \cdot r] = \Pr[B(5, p_0) < 5r]$$

ya que, si la proporción poblacional es p_0, $Z \mapsto B(5, p_0)$.

Pero, la probabilidad de rechazar la hipótesis si es cierta es $\Pr_{p_0}[\hat{p} < r]$ para $p_0 \geq 0{,}5$, por lo que la condición a imponer resulta $\Pr_{p_0}[\hat{p} < r] \leq 0{,}05$, si $p_0 \geq 0{,}5$.

Se comprueba que esta condición se verifica si se cumple para el "peor" punto de la hipótesis, es decir, el punto que separa hipótesis y alternativa. En nuestro caso, este punto es $p_0 = 0{,}5$, y se comprueba que esa probabilidad es menor cuanto más pequeño es p_0. Basta, por tanto, con imponer la condición

$$\Pr_{0,5}[\hat{p} < r] = \Pr[B(5, 0{,}5) < 5 \cdot r] \leq 0{,}05$$

condición que, con ayuda de las tablas de la distribución binomial, nos proporciona como solución más ajustada $5 \cdot r = 0{,}5$, ya que para ese valor, la probabilidad es 0,0312. Por tanto, $r = \frac{0{,}5}{5} = 0{,}1$, y la región crítica resultará $C = \{\hat{p} < 0{,}1\}$ y, como la proporción muestral sólo toma los valores 0, 0,2, 0,4, 0,6, 0,8 y 1, esta región no es sino

$$C = \{\hat{p} = 0\}.$$

En definitiva, si se quiere correr un riesgo no superior al 5% de rechazar la hipótesis cuando es cierta, sólo se rechazará la hipótesis si ninguno de los cinco encuestados ha repetido. El riesgo efectivo, de hecho, no rebasa el 3,12%. Como puede verse, la reducción del riesgo implica la reducción de la región crítica, como es natural.

Con los datos de nuestra encuesta, la proporción muestral era $\hat{p} = 0{,}4$, con lo cual no rechazamos la hipótesis.

De forma análoga se construyen test de contraste de otro tipo de hipótesis para la proporción poblacional. Sintetizamos los resultados a continuación.

Contraste de proporciones

Llamemos p a la proporción poblacional, y $\hat{p}$ a la proporción muestral, obtenida a partir de una muestra de tamaño n. Sea α un nivel de significación, $0<\alpha<1$.

Hip. y alt.	*R. crítica*	*Valor crítico*
$H_0: p \le p_0$ $H_0: p > p_0$	$C=\{\hat{p}>r\}$	$\Pr[B(n,p_0)>n\cdot r]\le\alpha$
$H_0: p \ge p_0$ $H_0: p < p_0$	$C=\{\hat{p}<r\}$	$\Pr[B(n,p_0)<n\cdot r]\le\alpha$
$H_0: p = p_0$ $H_0: p \ne p_0$	$A=\{r_1<\hat{p}<r_2\}$	$\Pr[B(n,p_0)\le n\cdot r_1]\le\alpha/2$ $\Pr[B(n,p_0)\ge n\cdot r_2]\le\alpha/2$

Como puede verse, el segundo caso es el que hemos desarrollado en el caso anterior. El primero es su simétrico, y se construye de forma análoga. En cuanto al tercero, la idea no es muy diferente. Puesto que la hipótesis postula que la proporción poblacional es una concreta, p_0, se rechazará si la proporción muestral es muy pequeña ($\hat{p}\le r_1$) o muy grande ($\hat{p}\ge r_2$). Puesto que $\Pr_{p_0}[C]\le\alpha$, esta probabilidad se distribuye por igual a la izquierda de r_1 y a la derecha de r_2.

Las tablas de la distribución binomial suministran probabilidades para $n\le 20$. Si $n>20$ se utiliza la aproximación normal, bien de Z o bien, directamente, de $\hat{p}$. En efecto, si $n>20$,

$$Z \mapsto B(n,p) \approx N\left(n\cdot p, \sqrt{npq}\right),$$

donde $q=1-p$, y p es la proporción poblacional.

En consecuencia,

$$\hat{p}=\frac{Z}{n}\approx N\left(p,\sqrt{\frac{pq}{n}}\right).$$

Veamos el siguiente ejercicio.

Ejemplo 1. El ayuntamiento de Peralillos estudia poner una línea de autobús nocturna (*El murciélago*) que recoja a sus vecinos los fines de semana de los bares de Rocafiel, la cabecera de comarca. La línea se pondrá en marcha si más de 1.000 de los 3.000 vecinos piensan utilizarla. Para informarse, realiza una encuesta a 100 vecinos, de los que 30 afirman que usarán el servicio de autobús. Contrasta la afirmación de que el servicio no se pondrá finalmente en marcha, al nivel de significación $\alpha=0{,}05$.

Solución. La hipótesis a contrastar es $H_0 : p \leq \frac{1}{3}$, y la alternativa, $H_1 : p > \frac{1}{3}$. La región crítica será de la forma $C = \{\hat{p} > r\}$. El valor r se obtiene mediante la condición

$$\Pr_{p=\frac{1}{3}}[\hat{p} > r] \leq 0{,}05.$$

Ahora bien, como $n = 100 > 20$, si $p = \frac{1}{3}$,

$$\hat{p} \approx N\left(\frac{1}{3}, \sqrt{\left(\frac{1}{3}\cdot\frac{2}{3}\right)/100}\right), \text{ o sea}$$

$$\hat{p} \approx N\left(\frac{1}{3}, \frac{\sqrt{2}}{30}\right).$$

Por tanto,

$$\Pr_{p=\frac{1}{3}}[\hat{p} > r] = \Pr_{p=\frac{1}{3}}\left[\frac{\hat{p}-\frac{1}{3}}{\sqrt{2}/30} > \frac{r-\frac{1}{3}}{\sqrt{2}/30}\right] = \Pr\left[N(0,1) > \frac{r-\frac{1}{3}}{\sqrt{2}/30}\right].$$

La condición

$$\Pr\left[N(0,1) > \frac{r-\frac{1}{3}}{\sqrt{2}/30}\right] = 0{,}05.$$

suministra el valor $\frac{r-\frac{1}{3}}{\sqrt{2}}\cdot 30=1{,}64$, y por tanto, $r=0{,}41$.

La región crítica es, por tanto,

$$C=\{\hat{p}>0{,}41\}$$

De la encuesta a 100 vecinos, 30 piensan usar el servicio, luego $\hat{p}=0{,}3$, que no pertenece a la región crítica. No se puede, pues, rechazar la hipótesis de que el servicio nocturno no se pondrá en marcha.

2.1. Otro tipo de error

En los desarrollos anteriores hemos intentado limitar la proporción de resultados muestrales para los que se rechazaba la hipótesis cuando era cierta, a una cantidad, α, denominada nivel de significación. En general, ello conduce, como hemos visto, a reducir la región crítica, de forma que se rechace menos veces la hipótesis. Con sólo esta idea a la vista, no se entiende por qué no reducir más aun el valor de α, incluso tomar $\alpha=0$ (entonces, en general, la región crítica resultará vacía, nunca se rechazará la hipótesis).

El asunto es que no debe perderse de vista otro error que puede cometerse al realizar un test de contraste de hipótesis: la aceptación de la hipótesis cuando es falsa. Este error se denomina **de tipo II**, para distinguirlo del anterior (rechazar la hipótesis cuando es cierta), que se denomina **error de tipo I**.

Ejemplo 2. Para el ejemplo 1, calcúlese la probabilidad del error de tipo II, esto es, la probabilidad de que se acepte $H_0: p\leq\frac{1}{3}$ siendo falsa.

Solución. Si la proporción poblacional es p, la probabilidad de aceptar la hipótesis es la probabilidad de la región de aceptación,

$$\Pr_p[A] = \Pr_p[\hat{p} \le 0{,}41]$$

Por tanto, la probabilidad de aceptarla si es falsa será la probabilidad anterior, cuando $p > \frac{1}{3}$.

Por ejemplo, si $p = 0{,}5$, $\hat{p} \approx N\left(0{,}5, \sqrt{\frac{0{,}5 \cdot 0{,}5}{100}}\right)$, o sea, $\hat{p} \approx N(0{,}5\ ,\ 0{,}05)$, de donde

$$\begin{aligned} \Pr_{p=0,5}[A] &= \Pr_{p=0,5}[\hat{p} \le 0{,}41] = \Pr_{p=0,5}\left[\frac{\hat{p} - 0{,}5}{0{,}05} \le \frac{0{,}41 - 0{,}5}{0{,}05}\right] = \\ &= \Pr[N(0,1) \le -1{,}8] = 1 - \Pr[N(0,1) \le 1{,}8] = \\ &= 1 - 0{,}9641 = 0{,}0359 \end{aligned}$$

Si $p = 0{,}8$, análogamente, $\hat{p} \approx N\left(0{,}8\ ,\ \frac{4\sqrt{2}}{100}\right)$, por lo que

$$\Pr_{p=0,8}[A] = \Pr\left[N(0,1) \le \frac{100(0{,}41 - 0{,}8)}{4\sqrt{2}}\right] = \Pr[N(0,1) \le -6{,}89] = 0{,}0000$$

Un último cálculo; si $p = 0{,}34$,

$$\Pr_{p=0,34}[A] = \Pr\left[N(0,1) \le \frac{100(0{,}41 - 0{,}34)}{0{,}0474}\right] = \Pr[N(0,1) \le 1{,}48] = 0{,}93$$

En definitiva, el test que obtuvimos en el ejercicio 1 sólo rechaza hipótesis ciertas con probabilidades no superiores a 0,05, pero puede aceptar hipótesis falsas con probabilidades que dependen del valor de p, pero que pueden ser muy grandes. Por ejemplo, el último cálculo indica que si $p = 0{,}34$ (y, por tanto, la hipótesis es falsa), la aceptaríamos para un 93% de los resultados muestrales.

Nótese que, a lo largo de los ejemplos anteriores, cuando el resultado muestral no estaba en la región crítica, decíamos que la hipótesis *no se rechazaba*, en lugar de decir que se aceptaba. Puede

ahora comprenderse el motivo de esta forma de hablar. Al fijarnos únicamente en los errores de tipo I, pocas veces se rechazan hipótesis ciertas, pero muchas se aceptan hipótesis falsas. La conclusión, pues, es que el no rechazo de la hipótesis no puede traducirse automáticamente por la aceptación, salvo con grandes riesgos de error.

Esta idea implica también que debe tomarse el nivel de significación tan alto como sea posible, ya que entonces aumentará la región crítica, disminuirá la de aceptación, y disminuirá asimismo la probabilidad de aceptar hipótesis falsas. Nótese también que debe intentarse apurar el nivel de significación, esto es, que la acotación que establece que la probabilidad de error de primer tipo debe ser **menor o igual** que el nivel de significación, debe estar lo más ajustada posible, llegando incluso a la igualdad.

2.2. Muestreo sin reemplazamiento

Como es sabido, la distribución binomial estudia las probabilidades de obtener un cierto número de veces un suceso (un cierto número de *éxitos*) cuando se repite n veces un experimento, siempre que estas repeticiones sean independientes unas de otras. En el primero de los test anteriores, aceptábamos que el número de encuestados que habían perdido un curso (el suceso "éxito" sería aquí que un alumno hubiera perdido un curso) seguía una distribución binomial, $B(5, p)$, donde p era la proporción poblacional, es decir, la probabilidad de "éxito" en una tirada.

Este resultado está ligado al hecho de que cada encuestado puede serlo más de una vez, o sea, que cada encuesta se escoge del conjunto de la población.

En la clase había 10 alumnos que no habían perdido curso, y 20 que sí lo habían perdido. Cuando se escoge el primer encuestado, la proporción poblacional es $p = \frac{20}{30} = \frac{2}{3}$. Al elegir el segundo, si en el sorteo participan los 30 encuestados, la proporción poblacional vuelve a ser de dos tercios; pero si no participa el que ya ha sido

encuestado, la proporción poblacional cambia. Si el primer encuestado ha perdido algún curso, la nueva proporción poblacional será $\frac{19}{29}$, y si no lo ha perdido, de $\frac{20}{39}$.

En consecuencia, si no hay reemplazamiento (un encuestado no participa en los siguientes sorteos, o, lo que es lo mismo, los encuestados se eligen simultáneamente) los resultados de cada encuesta no son independientes, y el número total de éxitos (numerador de la proporción muestral) no seguirá una distribución binomial.

No obstante, si el número de encuestados es pequeño en comparación con el total de la población, la proporción no cambia prácticamente, por lo que, aunque no se encueste varias veces al mismo individuo, puede trabajarse como si ocurriera. Se acepta que la **fracción de muestreo** (cociente entre número de encuestados y tamaño de la población) no debe rebasar 0,1 (el 10%) para que la aproximación sea válida.

2.3. Un problema normal

Ejemplo 3. En una reunión de directores de Institutos, el del "Emilio Porsche" se jacta de que la nota media de sus alumnos es, al menos, de 6, la superior en Valladolid. La directora del "Leopoldo Calvo" cree que no es cierto, y envía a varios alumnos al primero de los institutos para obtener varios expedientes. Los alumnos consiguen seis de ellos al azar del archivador en que se guardan, que arrojan las calificaciones 6,1, 2,2, 7,3, 4,4, 3,5 y 5,3. Contrasta la afirmación del director del "Emilio Porsche" a un nivel de significación de 0,1, suponiendo que las calificaciones siguen una distribución normal.

Solución. Si llamamos μ (se lee "mu") a la nota media del primer Instituto, la hipótesis a contrastar es $H_0 : \mu \geq 6$, y la alternativa, $H_1 : \mu < 6$. Si llamamos X a las calificaciones de sus alumnos sabemos que $X \mapsto N(\mu, \sigma)$.

Cuando estudiamos los problemas de estimación, vimos que una buena estimación de la media poblacional, μ, era la media muestral, $\bar{x}=\frac{x_1+\cdots+x_n}{n}$. Con las ideas que estudiamos en los contrastes de proporciones, parece razonable rechazar la hipótesis para pequeños valores de la media muestral, ya que la alternativa especifica pequeños valores para la media poblacional. Tomemos entonces

$$C=\{\bar{x}<r\}$$

como región crítica. Puesto que el nivel de significación es $\alpha=0{,}1$, debe imponerse la condición

$$p_\mu[C]=p_\mu[\bar{X}<r]\leq 0{,}1 \text{ , para } \mu\geq 6$$

donde p_μ indica *probabilidad, cuando la media poblacional es* μ. Se impone, por tanto, la condición de que para no más del 10% de las realizaciones muestrales se rechace la hipótesis cuando es cierta.

Dos cuestiones debe aceptar el alumno. La primera, que ya utilizamos al hablar de estimación, indica que si la población de notas, X, sigue una distribución normal $N(\mu,\sigma)$, las medias muestrales para una muestra de tamaño n siguen también una distribución normal, $N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$.

La segunda, que la probabilidad de error de primer tipo anterior alcanza su máximo para el "peor" punto de la hipótesis, esto es, para $\mu=6$; o sea, que

$$p_\mu[C]\leq p_{0,6}[C], \text{ para } \mu\geq 6$$

De hecho, lo que ocurre es que la probabilidad de la región crítica anterior es una función decreciente de μ. Esta segunda cuestión implica que basta con imponer la condición

$$p_{0,6}[C]=0{,}1$$

para que la probabilidad de error de primer tipo no rebase el nivel de significación.

Pero esta última condición permite obtener el valor de r. En efecto,

$$p_{0,6}[C]=p_{0,6}[\overline{X}<r]=p_{0,6}\left[\frac{\overline{X}-0,6}{\sigma/\sqrt{n}}<\frac{r-0,6}{\sigma/\sqrt{n}}\right]$$

Ahora bien, si $\mu=0,6$, $\overline{X}\mapsto N\left(0,6,\dfrac{\sigma}{\sqrt{n}}\right)$, por lo que

$$\frac{\overline{X}-0,6}{\sigma/\sqrt{n}}\mapsto N(0,1)$$

En definitiva, la ecuación anterior resulta

$$p_{0,6}[C]=p\left[N(0,1)<\frac{r-0,6}{\sigma/\sqrt{n}}\right]=0,1$$

lo que, con las tablas de la distribución normal, permite obtener

$$\frac{r-0,6}{\sigma/\sqrt{n}}=-1,29$$

y, en consecuencia

$$r=0,6-1,29\cdot\frac{\sigma}{\sqrt{n}}.$$

Hasta ahora no hemos dedicado atención al valor de la desviación típica poblacional, σ, pero el resultado anterior no resulta útil si esta desviación es desconocida. Si, por ejemplo, $\sigma=1,5$,

$$r=0,6-1,29\cdot\frac{1,5}{\sqrt{6}}=0,6-0,79=-0,19.$$

Podemos entonces escribir la región crítica,

$$C=\{\bar{x}<-0{,}19\}$$

Para concluir, calculemos el valor de la media muestral para nuestros 6 encuestados,

$$\bar{x}=\frac{6{,}1+2{,}2+7{,}3+4{,}4+3{,}5+5{,}3}{6}=4{,}8$$

El resultado muestral no pertenece, por tanto, a la región crítica, por lo que no se rechaza la hipótesis de que la nota media del Emilio Porsche es de seis o superior.

El resultado anterior parecerá, sin duda alguna, sorprendente. Para rechazar la hipótesis debiéramos obtener una media muestral de notas inferior a -0,19 puntos. Ello indica que nunca rechazaremos la hipótesis, por lo que parece dudoso que la probabilidad de rechazarla si es cierta no sea igual a cero, sino a 0,1. Nótense, con todo, las siguientes cuestiones:

1. Estamos suponiendo que las calificaciones, X, siguen una distribución normal, luego estamos dispuestos, formalmente, a aceptar calificaciones negativas. La costumbre (perniciosa) de suponer que una variable no negativa sigue una distribución normal, aun siendo muy ventajosa desde el punto de vista analítico, crea estas aparentes paradojas.

2. Utilizamos muy poca información (sólo encuestamos a 6 alumnos) y queremos tener un riesgo pequeño (si la hipótesis es cierta, únicamente un 10% de resultados muestrales nos conducen a aceptarla), cuando las calificaciones están bastante dispersas ($\sigma=1{,}5$). Puede observarse en la expresión que proporciona r que el número que se resta a 0,6 cambiaría:

   - modificando α. Si, por ejemplo, $\alpha=0{,}4$, obtendríamos -0,27, en lugar de -1,29.

- cambiando el tamaño de la muestra. Si encuestamos a 100 alumnos en lugar de los 6 elegidos, obtendríamos $r=0{,}4$ para $\alpha=0{,}1$.

Dejemos, con todo, estas discusiones. Sistematicemos los resultados obtenidos, incluyendo además la forma de proceder cuando la hipótesis y la alternativa tienen un aspecto distinto del propuesto en el ejercicio anterior.

Contraste de medias de poblaciones normales (σ conocida)

Sea X una población normal, $N(\mu,\sigma)$, con σ conocida, y sea $\hat{\mu}$ la estimación de μ que resulta de calcular la media muestral para una muestra de tamaño n,

$$\hat{\mu}=\bar{x}=\frac{x_1+\cdots+x_n}{n}.$$

Sea α un nivel de significación, $0<\alpha<1$.

Hip. y alt.	*R. crítica*	*Valor crítico*
$H_0:\mu\le\mu_0$ $H_1:\mu>\mu_0$	$C=\{\hat{\mu}>r\}$	$\Pr[N(0,1)>r'\,]\le\alpha$
$H_0:\mu\ge\mu_0$ $H_1:\mu<\mu_0$	$C=\{\hat{\mu}<r\}$	$\Pr[N(0,1)<r']\le\alpha$
$H_0:\mu=\mu_0$ $H_1:\mu\ne\mu_0$	$A=\{r_1<\hat{\mu}<r_2\}$	$\Pr[N(0,1)\le r_1']\le\alpha/2$ $\Pr[N(0,1)\ge r_2']\le\alpha/2$

siendo $r'=\dfrac{r-\mu_0}{\sigma}\sqrt{n}$, esto es, $r=\mu_0+r'\dfrac{\sigma}{\sqrt{n}}$, y lo mismo para r_1 y r_2.

Cuando σ es desconocida, pueden realizarse los mismos contrastes sin excesivas modificaciones. Se sustituye σ por una estimación suya, la cuasi-desviación típica muestral, s_c, raíz cuadrada positiva de la cuasi-varianza muestral, s_c^2, definida como

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

donde s^2 es la varianza del resultado muestral. Este cambio obliga, asimismo, a modificar la búsqueda del valor crítico, que ahora se encontrará en las tablas de la denominada t de *Student* con $n-1$ grados de libertad, abreviadamente t_{n-1}. El resultado general es muy similar al anterior, aunque lo incluimos para que pueda servir de referencia.

Sea X una población normal, $N(\mu,\sigma)$, con σ desconocida, y sea $\hat{\mu}$ la estimación de μ que resulta de calcular la media muestral para una muestra de tamaño n,

$$\hat{\mu} = \bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Sea α un nivel de significación, $0 < \alpha < 1$.

Hip. y alt.	*R. crítica*	*Valor crítico*
$H_0 : \mu \le \mu_0$ $H_1 : \mu > \mu_0$	$C = \{\hat{\mu} > r\}$	$\Pr[t_{n-1} > r' \,] \le \alpha$
$H_0 : \mu \ge \mu_0$ $H_1 : \mu < \mu_0$	$C = \{\hat{\mu} < r\}$	$\Pr[t_{n-1} < r'] \le \alpha$
$H_0 : \mu = \mu_0$ $H_1 : \mu \ne \mu_0$	$A = \{r_1 < \hat{\mu} < r_2\}$	$\Pr[t_{n-1} \le r_1'] \le \alpha/2$ $\Pr[t_{n-1} \ge r_2'] \le \alpha/2$

siendo $r' = \dfrac{r-\mu_0}{s_c}\sqrt{n}$, esto es, $r = \mu_0 + r'\dfrac{s_c}{\sqrt{n}}$, y lo mismo para r_1 y r_2.

2.4. Ejercicios propuestos

Ejercicio 6.1. *Se quiere contrastar que la proporción de varones nacidos en un hospital es de 0,5 (o sea, el 50%), frente a la alternativa de que esa proporción es distinta del 50%.*

a) En un mes se escogen por sorteo 20 nacimientos, de los que 13 resultan ser niñas. Construir un test para realizar el contraste al nivel de significación $\alpha = 0{,}05$.

b) De los nacidos en el mes siguiente, se seleccionan otros 10 nacimientos, resultando 6 niñas. Contrastar nuevamente la anterior afirmación con la muestra de 30 nacimientos total.

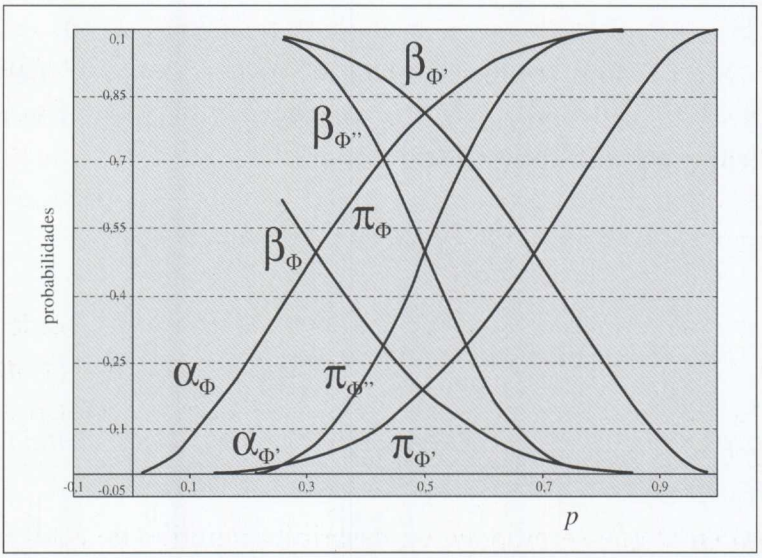

Figura 6.1

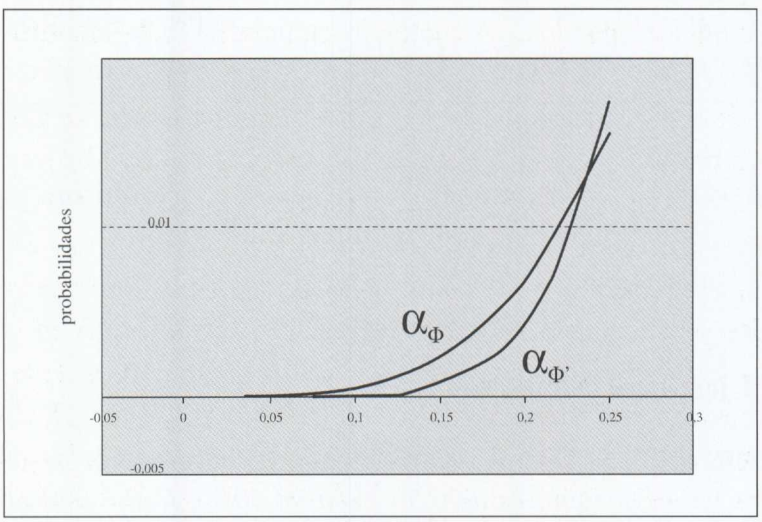

Figura 6.2

BIBLIOGRAFÍA COMENTADA

Como indicamos en la Introducción, la teoría de los contrastes de hipótesis no es, en general, conocida por los profesores de Matemáticas en Bachillerato. Además, la novedad de su notación y de sus motivaciones hace que no se pueda considerar una prolongación intuitiva del Cálculo de probabilidades o de la teoría de la estimación. Este es el motivo por el que hemos preferido desarrollar un material orientado a la formación del profesorado, en lugar de materiales didácticos para los alumnos, que sólo se presentan como propuestas finales.

Por el mismo motivo, hemos construido un material que puede considerarse autosuficiente para el profesor, al menos como un primer contacto, de forma que pueda seguirse sin recurrir a otros recursos bibliográficos, en general más densos y áridos.

Un libro en el que se propone un desarrollo intuitivo de estos materiales es el ya citado de Llopis (LLOPIS PÉREZ, J. *La estadística: una orquesta hecha instrumento*. Ariel Ciencia. Barcelona, 1996), siendo posiblemente la primera lectura a recomendar cuando el profesor se adentra por los vericuetos inferenciales de la Estadística.

Referencias históricas sobre el tema no son abundantes, ya que esta teoría se ha desarrollado en un momento en que los estadísticos eran ya precisamente eso, estadísticos, y no como en otras épocas, en las que personajes como Galileo, físico en el más amplio sentido de la palabra, se interesaba por problemas como los del cálculo de probabilidades (chances) para los juegos de dados, recibiendo en consecuencia una mayor atención por parte de los historiadores de la ciencia. Para estas referencias recomendamos la MacTutor History of Mathematics (en la Web de la Escuela de Matemáticas y Estadística de la universidad escocesa de St. Andrews, **http://www-gap.dcs.st-and.ac.uk/~history/index.html**)

Desde el punto de vista de los materiales escritos, existen algunos manuales de nivel accesible y contenido suficientemente completo, que seguramente colmarán las aspiraciones de quienes deseen una mayor formalización de los resultados, a saber, los libros de Casas (CASAS, J. *Inferencia Estadística para economía y administración de empresas*. Centro de Estudios Ramón. Areces. Madrid, 1996), Ruiz Maya y Martín Pliego (MARTÍN PLIEGO, F. J. Y RUIZ MAYA, L. *Estadística I: Probabilidad*. AC. Madrid, 1995) y Peña (PEÑA, D. *Fundamentos de Estadística*. Alianza Editorial. Madrid, 2001). Quienes lean con cierta soltura el inglés científico encontrarán muy interesante el manual de Rohatgi (ROHATGI, V. K. *Statistical Inference*. Wiley. New York, 1994) si bien resulta difícil de encontrar y es de un precio ciertamente elevado. Posee la ventaja este libro de estudiar conjuntamente las cuestiones inferenciales (estimación y contrastes) para cada una de las distribuciones tipo, concentrando los resultados referentes a la inferencia para las poblaciones normales o para las binomiales (proporciones)

Qué duda cabe que puede también recurrirse a los libros más clásicos, como Cramér (CRAMÉR, H. *Métodos matemáticos de Estadística*. Aguilar. Madrid, 1967), capítulos XXXI y XXXV, Mood y Graybill (MOOD, A. M. Y GRAYBILL, F. A. *Introducción a la Teoría de la Estadística*. Aguilar. Madrid, 1972), capítulo 12 o Fourgeaud y Fuchs (FOURGEAUD, C. Y FUCHS, A. *Statistique*. Dunod. París, 1967), capítulos 16 a 19. Pero estos libros, junto con el de Lehmann (LEHMAN, E.L. *Testing statistical hypotheses*. John Wiley & Sons. New York, 1986) en su conjunto, sólo se dirigen a lectores muy especializados y son de muy poca ayuda para la docencia estadística en los ámbitos de las enseñanzas medias.

ANEXO
Experimentación en el aula

Como señalamos en la Introducción, las experiencias realizadas con los estos materiales lo fueron de forma sistemática (y, por tanto, con un suficiente grado de evaluación) en el momento en que el proyecto fue finalizado, hace ahora más de 6 años. Desde entonces, los materiales han sido empleados en el aula de forma dispersa, recibiendo aportaciones individuales que no han sido incorporadas porque su diferente formato impediría una presentación organizada.

Por otro lado, en el tiempo transcurrido han aparecido y se han difundido otros medios de actuación en el aula, que en este trabajo no se utilizan. También se ha modificado la trayectoria en el proceso de aprendizaje de los alumnos, de sus aptitudes y actitudes, no sólo en su actividad dentro de los centros educativos, sino en el conjunto de su aprendizaje vital, modificación que ha reforzado los elementos visuales y dinámicos frente a los estáticos, mermando su capacidad de comprensión de los mensajes escritos.

Por todo ello, el resumen de la experimentación será necesariamente breve, señalando únicamente las características y conclusiones que hoy se mantienen vigentes por depender de los contenidos o de su articulación temporal, y no de las técnicas expositivas.

El **objetivo último** de dicha experimentación consiste en la validación, en su caso, de las propuestas de desarrollo curricular formuladas. Más en concreto, supone la valoración del grado en que los alumnos progresan, con la ayuda de los materiales, en la adquisición de las siguientes **capacidades**:

- Conseguir una familiarización con el lenguaje estadístico

- Facilitar y propiciar la obtención de información procedente de diversos contextos por parte de los propios alumnos

- Integrar las nuevas tecnologías de lá información en la educación, lo que en su día se interpretaba como el tratamiento informático de los datos recogidos, a través de una hoja de cálculo u otros programas de cálculo estadístico.

- Desarrollar en los alumnos la actitud crítica frente a la información recibida en forma estadística.

- Ayudar a los alumnos a establecer relaciones entre las matemáticas y el entorno social, cultural y económico, para que reconozcan el importante papel que aquéllas desempeñan en nuestra cultura.

- Utilizar los conocimientos matemáticos adquiridos para interpretar críticamente los mensajes, datos e informaciones que aparecen en los medios de comunicación y otros ámbitos sobre cuestiones económicas y sociales de la actualidad.

- Destacar los aspectos transversales de esta disciplina mediante la ocupación activa en problemas relacionados con otros contenidos del currículo, y

- Adiestrar a los alumnos en la utilización de técnicas estadísticas elementales para tomar decisiones en situaciones que se ajusten a modelos de probabilidad conocidos.

Los **contenidos** de los materiales curriculares empleados para Bachillerato se transcriben en este volumen, por lo que no insistiremos en ellos. No obstante, un breve resumen (**secuenciación y temporalización**) ayudará a estructurar los materiales experimentales, incluyendo nuestra valoración del tiempo empleado (secuenciación) en el desarrollo de los contenidos (debe tenerse en cuenta que este tiempo es el utilizado en su día. Los perfiles curriculares previos de

los alumnos y la programación han cambiado desde entonces, por lo que la validez es únicamente relativa):

- ***Primer curso***: Se consolidan, con un mayor grado de formalización, los conocimientos descriptivos univariantes que ya forman parte de los contenidos de la E.S.O. Se abordan los contenidos referentes a la tabulación y resumen de datos cuando una población se estudia según dos características. Se construyen e interpretan coeficientes de correlación lineal, poniendo en relación este concepto con el de causalidad. Se introduce al alumno en la noción de variable aleatoria, y en los cálculos probabilísticos que implica la utilización de estos modelos numéricos. Se proponen, desde un contexto aplicado y práctico, las distribuciones binomial y normal, junto con sus características y propiedades más relevantes y la relación entre ellas. La ***duración*** puede establecerse en 4 semanas lectivas.

- ***Segundo curso***: En el Bachillerato de Ciencias Sociales, único en que este currículum se imparte, se profundiza en el concepto de probabilidad y en las técnicas de asignación de probabilidades en experimentos aleatorios, incluidas las situaciones en que el experimento tiene una naturaleza compuesta desde una perspectiva temporal. Se enseña a identificar situaciones en las que deben aplicarse el Teorema de probabilidad total y el teorema de Bayes y se aplica a situaciones concretas, profundizando en la noción de dependencia e independencia estocástica. Se enseña a conocer las etapas de un estudio estadístico: elección de una muestra, estudio de la representatividad de la misma e inferencia de conclusiones sobre alguna característica de la población (medias y proporciones). Se enseña a enunciar hipótesis estadísticas para medias normales y proporciones, elaborando algún contraste de hipótesis paramétricas. La ***duración*** se establece en 4 semanas lectivas.

Valoración de la experiencia en el aula: Tanto en el I.E.S. Emilio Ferrari como en el I.E.S. Leopoldo Cano, se ha experimenta-

do con los materiales a lo largo de los cursos 1996-97 y 1997-98. En general, los resultados pueden considerarse satisfactorios, especialmente en dos aspectos:

- La calidad de los materiales teóricos, que han permitido a los alumnos recoger las líneas maestras del discurso que lleva desde la información hasta la elaboración estadística.

- Los ejercicios relacionados con datos referentes a la misma comunidad educativa o, más aún, a los propios alumnos suponen una motivación extra que mantiene la atención del alumno y su interés por obtener resultados no evidentes (encuestas en el aula o en cursos próximos, datos familiares, calificaciones)

Probablemente, las mayores dificultades se derivan de la insuficiente preparación matemática de base por parte de los alumnos, especialmente en los materiales correspondientes a la probabilidad y a la inferencia.

Digamos finalmente, que ha resultado algo escaso el material correspondiente a muestreo (tipos de muestreo, distribución de medias y proporciones) así como a los contrastes de hipótesis, si bien en este último caso recordemos que el material que se presenta en este libro está fundamentalmente dirigido a la preparación del profesor, teniendo las propuestas didácticas una presencia puramente anecdótica.

Una experiencia didáctica con el profesorado

Los temas correspondientes a la Inferencia estadística (capítulos 4, 5 y 6 de este volumen), especialmente el tema 6 (Test de contraste de hipótesis), están pensados, no como unidades didácticas destinadas a los alumnos del Bachillerato de Ciencias Sociales, sino a la formación del profesorado de Matemáticas implicado en dichas enseñanzas. Así, los profesores del equipo han utilizado este trabajo más para la preparación de sus clases que para la secuenciación de sus enseñanzas.

Los materiales inferenciales de estos tres últimos capítulos han sido objeto de una experiencia en la formación del profesorado. En concreto, se utilizaron en el curso "Procedimientos de actuación en el aula de Matemáticas aplicadas a las CC.SS.II", impartido en noviembre y diciembre de 1999, dirigido por Tomás Ortega del Rincón y coordinado por Mª Ángeles Gil Blanco, asesora del CPR de Almazán y actualmente directora del Centro de Profesorado e Innovación educativa de Soria. Se trató de un curso de ámbito provincial, dentro del Programa de Actualización Científica y Didáctica del Plan provincial de Formación. En él participaron 15 profesores, de los que 14 finalizaron con derecho a certificación. Dentro del curso, estos materiales se emplearon en 4 sesiones de 3 horas, esto es, a lo largo de 12 horas del curso.

Extractemos las notas más relevantes del acta de evaluación. En cuanto a la valoración de los materiales y contenidos para su uso directo en el aula reciben puntuaciones (en una escala de 1 a 4 y, por tanto, sobre valores medios de 2,5) en torno a la media potencial. En concreto, la *Eficacia del material didáctico* se valora con un 2,7, al igual que la *Capacidad de actualización de los contenidos científicos* y la *Posibilidad de incorporar al aula nuevas tecnologías*. Una puntuación algo inferior recibe la *Capacidad para hacer más atractiva la asignatura* para los alumnos (2,1).

Superior valoración obtienen los capítulos inferenciales en su capacidad formativa del profesorado. Así, tanto el *Interés del contenido* como la *Estrategia metodológica* se valoran con 3,6, y con 3,3 el *Grado de utilidad para el trabajo* de los profesores asistentes. La *Adecuación de los recursos* se valora con 3,4. La puntuación global que los asistentes dan a la parcela del curso centrada en los materiales inferenciales resulta ser de 3,6.

BIBLIOGRAFÍA

AITCHISON, J., y BROWN, J. A. C. *The lognormal distribution.* Cambridge University Press. Cambridge (UK), 1957.

ANDERSON, I. *Introducción a la Combinatoria.* Vicens Vives. Barcelona, 1993.

ARANDA, J. y otros. *Introducción a la Estadística Económica y Empresarial. Ejercicios.* DM y PPU, Murcia y Barcelona, 1994.

ARDANUY, R., y SOLDEVILLA, M. M. *Estadística Básica,* Hespérides. Salamanca, 1992.

ARNÁIZ, G. *Introducción a la Estadística Teórica.* Lex Nova. Valladolid, 1986.

AYUSO, M.; CORRALES, H.; GUILLÉN, M.; PÉREZ-MARÍN, A. M., y ROJO, J. L. *Estadística actuarial Vida.* Ed. Universitat de Barcelona, UB51 manuales. Barcelona, 2001.

CANAVOS, G. C. *Probabilidad y Estadística: Aplicaciones y Métodos.* McGraw Hill. México, 1992.

CASAS, J. *Inferencia estadística para economía y administración de empresas.* Centro de Estudios Ramón Areces. Madrid, 1996.

CASAS, J., y SANTOS, J. *Introducción a la estadística para economía y administración de empresas.* Centro de Estudios Ramón Areces. Madrid, 1995.

CHAO, L. L. *Estadística para las ciencias administrativas.* McGraw Hill. 3.ª ed. Bogotá, 1993.

CLAIRIN, R., y BRION, P. H. *Manual de muestreo.* La Muralla/ Hespérides. Madrid/Salamanca, 2001.

COLLS, S., y GUIJARRO, M. *Estadística aplicada a la historia y las ciencias sociales.* Pirámide. Madrid, 1998.

CRAMÉR, H. *Elementos de la teoría de probabilidades y algunas de sus aplicaciones.* Aguilar. Madrid, 1977.

CRAMÉR, H. *Métodos matemáticos de Estadística.* Aguilar. Madrid, 1967.

CUADRAS, C. M.; ECHEVARRÍA, B.; MATEO, J., y SÁNCHEZ, P. *Fundamentos de estadística: Aplicación a las ciencias humanas.* P.P.U. Barcelona, 1984.

DAW, R. H., y PEARSON, E. S. *"Studies in the history of probability and statistics. XXX: Abraham de Moivre's 1733 derivation of the normal curve: a bibliographical note". Biométrica,* 59. 1972. Págs. 677-680.

DURÁ. J. M., y LÓPEZ, J. M. *Fundamentos de Estadística: Estadística descriptiva y modelos probabilísticos para la inferencia.* Ariel Economía. Barcelona, 1988.

ESCUDER, R., y MURGUI, J. S. *Estadística aplicada. Economía y Ciencias Sociales.* Tirant lo Blanch. Valencia, 1995.

FELLER, W. *Introducción a la teoría de probabilidades y sus aplicaciones.* Vol. I y II. Limusa. México, 1975 y 1978.

FERNÁNDEZ, C., y FUENTES, F. *Curso de Estadística Descriptiva. Teoría y práctica.* Ariel Economía. Barcelona, 1995.

FERNÁNDEZ-ABASCAL, H.; GUIJARRO, M.; ROJO, J. L., y SANZ, J. A. *Cálculo de Probabilidades y Estadística.* Ariel Economía. Barcelona, 1994.

FERNÁNDEZ-ABASCAL, H.; GUIJARRO, M.; ROJO, J. L., y SANZ, J. A. *Ejercicios de cálculo de probabilidades: resueltos y comentados.* Ariel Matemática. Barcelona, 1995.

FERNÁNDEZ DE TROCÓNIZ, A. *Introducción a las teorías de las probabilidades. Estadística clásica y estadística bayesiana.* Autoeditado. Bilbao, 1980.

FISHER, R. A. *Statistical Methods, Experimental Design and Scientic Inference.* Editado por J. H. Bennett con un prólogo de F. Yates. Oxford University Press. Oxford, 1991.

FREIRÁ, M. y otros. *Análisis exploratorio de datos: nuevas técnicas estadísticas.* PPU. Barcelona, 1992.

FREUND, J. E., y SIMON, G. A. *Estadística Elemental.* 8.ª ed. Prentice-Hall. México, 1994.

FOURGEAUD, C., y FUCHS, A. *Statistique.* Dunod. París, 1967.

GARCÍA BARBANCHO, A. *Estadística teórica básica: Probabilidad y modelos probabilísticos.* Ariel Economía. Barcelona, 1992.

GNEDENKO, B. V. *The Theory of Probability.* MIR. Moscú, 1978.

GROOT, M. H. DE. *Probabilidad y Estadística.* Addison-Wesley Iberoamericana. México D.F., 1988.

HANKE, J. E., y REITSCH, A. G. *Estadística para negocios.* Irwin. Madrid, 1995.

HUXLEY, G. L. *"The mathematical work of Edmond Halley". Scripta Math,* 24, 1959, págs. 265-273.

JEFFREYS, H. *Theory of Probability.* 3.ª ed. (1.ª ed. 1939). Clarendon Press. Oxford, 1961.

JOHNSON, N., y KOTZ, S. *Distributions in Statistics. Vol. 1: Discrete Distributions.* Wiley. New York, 1969.

JOHNSON, N., y KOTZ, S. *Distributions in Statistics. Vol. 2: Continuous Univariate Distributions-1.* Wiley. New York, 1970.

JOUNSON, N., y KOTZ, S. *Distributions in Statistics. Vol. 3: Continuous Univariate Distributions-2.* Wiley. New York, 1971.

JOHNSON, N., y KOTZ, S. *Distributions in Statistics. Vol. 4: Continuous Multivariate Distributions.* Wiley. New York, 1972.

KENDALL, M. G. *Enciclopedia Internacional de las Ciencias Sociales.* Vol. 4. Aguilar. Madrid, 1979. Pág. 404.

KOLMOGOROV, A. N. *Foundations of the Theory of Probability.* Chelsea. New York, 1956.

LAMBALGEN, M. van. *"Randomness and foundations of probability: von Mises' axiomatisation of random sequences".* En *Statistics, probability and game theory.* IMS Lecture Notes. Hayward, California, 1996.

LAPLACE, P. S. de. (1774). *"Determiner le mimieu que l'on doit prendre entre trois observations données d'un même phénomene".* En *Memoires de Mathématique et Phisique presentées á l'Académie Royale des Sciences par divers Savans,* 6. 1774. Págs. 621-625.

LAPLACE, P. S. de. *Théorie analytique des probabilités* (3e édition, 1820). Existe una edición reciente en LAPLACE OEUVRES, Tome VII. De Editions Jacques Gabay. París, 1995.

LEHMAN, E. L. *Testing statistical hypotheses.* John Wiley & Sons. New York, 1986.

LEVINE, D. M.; RAMSEY, P. P., y BERENSON, M. L. *Business Statistics for Quality and Productivity.* Prentice-Hall. New Jersey, 1995.

LÓPEZ CACHERO, M. *Fundamentos y métodos de estadística.* Pirámide. Madrid, 1989.

LLOPIS PÉREZ, J. *La estadística: una orquesta hecha instrumento.* Ariel Ciencia. Barcelona, 1996.

MACTUTOR HISTORY OF MATHEMATICS (Web de la Mathematics and Statistics School, St Andrews University. http://www-gap.dcs.st-and.ac.uk/~history/index.html).

MARTÍN-GUZMÁN, M. P., y M. PLIEGO, F. J. *Curso básico de estadística económica.* AC. Madrid, 1989.

MARTÍN PLIEGO, F. J. *Curso práctico de estadística económica.* AC. Madrid, 1987.

MARTÍN PLIEGO, F. Javier. *Introducción a la estadística económica y empresarial (teoría y práctica).* AC. Madrid, 1994.

MARTÍN PLIEGO, F. J., y RUIZ MAYA, L. *Estadística I. Probabilidad.* AC. Madrid, 1995.

MASON, R. D., y LIND, D. A. *Estadística para administración y economía.* Alfaomega. México, 1992.

MONTIEL A. M.; RIUS, F., y BARÓN, F. J. *Elementos básicos de estadística económica y empresarial.* Prentice Hall. Madrid, 1996 y 1997.

MOOD, A. M., y GRAYBILL, F. A. *Introducción a la Teoría de la Estadística.* Aguilar. Madrid, 1972.

MOORE, D. S. *Estadística aplicada básica.* Antoni Bosch. Barcelona, 1995.

MORA CHARLES, Marisol de. *Los inicios de la Teoría de la Probabilidad. Siglos XVI y XVII.* Servicio editorial UPV. Bilbao, 1989.

NEWBOLD, P. *Estadística para los negocios y la economía.* Prentice Hall. Madrid, 1997.

NORTES, A. *Estadística teórica y aplicada.* DM y PPU. Murcia y Barcelona, 1991.

PEARSON, E. S., y KENDALL, M. G. *Studies in the History of Statistics and Probability.* Ed. Griffin. London, 1970.

PEÑA, D. *Estadística. Modelos y Métodos 1. Fundamentos.* Alianza Editorial. Madrid, 1991.

PEÑA, D. *Fundamentos de Estadística.* Alianza Editorial. Madrid, 2001.

PEÑA, Daniel, y ROMO, Juan. *Introducción a la estadística para las ciencias sociales.* McGraw-Hill. Madrid, 1997.

PÉREZ SUÁREZ, Rigoberto. *Análisis de datos económicos I. Métodos descriptivos.* Pirámide. Madrid, 1993.

PÉREZ, R., y LÓPEZ, A. J. *Análisis de datos económicos II. Métodos inferenciales.* Pirámide. Madrid, 1997.

RODRÍGUEZ, J., y ARENALES, C. *Problemas de estadística económica.* Pirámide. Madrid, 1988.

ROHATGI, V. K. *An Introducción to Probability Theory and Mathematical Statistics.* Wiley. New York, 1977.

ROHATGI, V. K. *Statistical Inference.* Wiley. New York, 1994.

RUIZ MAYA, L., y MARTÍN PLIEGO, F. J. *Estadística II: Inferencia,* AC. Madrid, 1995.

SANCHÍS, C. y otros. *Hacer estadística.* Alhambra Longman. Madrid, 1986.

SMITH, D. E. *"Legendre on least squares",* in *A source book of mathematics.* McGraw-Hill. New York, 1929. Reimpresión de Ed. Dover. New York, 1959.

STIGLER, S. M. *The History of Statistics. The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. Cambridge, Mass., 1986.

TANUR, J. M.; MOSTELLER, F.; KRUSKAL, W. H.; LEHMAN, E. L.; LINK, R. F.; PIETERS, R. S., y RISING, G. R. (Eds.). *La Estadística: Una guía de lo desconocido.* Alianza Ed. Madrid, 1992.

WALPOLE, R. E., y MYERS, R. H. *Probabilidad y Estadística.* McGraw Hill. 4.ª ed. México, 1991.

WONNACOTT, T. H., y WONNACOTT, R. J. *Introducción a la Estadística.* Limusa. México, 1979.

EDICIONES DEL INSTITUTO SUPERIOR
DE FORMACIÓN DEL PROFESORADO

Subdirección General de Información y Publicaciones
del Ministerio de Educación, Cultura y Deporte

EDICIONES DEL INSTITUTO SUPERIOR DE FORMACIÓN DEL PROFESORADO

Subdirección General de Información y Publicaciones del Ministerio de Educación, Cultura y Deporte

El Instituto Superior de Formación del Profesorado tiene como objetivo impulsar, incentivar, financiar, apoyar y promover acciones formativas realizadas por las instituciones, Universidades y entidades sin ánimo de lucro, de interés para los docentes de todo el Estado Español que ejercen sus funciones en las distintas Comunidades y Ciudades Autónomas. Pero, tan importante como ello, es difundir, extender y dar a conocer, en el mayor número de foros posible, y al mayor número de profesores, el desarrollo de estas acciones. Para cumplir este objetivo, el I.S.F.P. pondrá a disposición del profesorado español, con destino a las bibliotecas de Centros y Departamentos, **dos colecciones**, divididas cada una en cuatro series.

Con estas colecciones, como acabamos de señalar, se pretende difundir los contenidos de los cursos, congresos, investigaciones y actividades que se impulsan desde el Instituto Superior de Formación del Profesorado, con el fin de que su penetración difusora en el mundo educativo llegue al máximo posible, estableciéndose así una fructífera intercomunicación dentro de todo el territorio del Estado.

La primera de nuestras colecciones se denomina **Aulas de Verano**, y pretende que todo el profesorado pueda acceder al conocimiento de las conferencias, ponencias, mesas redondas, talleres y actividades profesionales docentes que se desarrollan durante los veranos en la *Universidad Internacional Menéndez Pelayo de Santander*, en los cursos de la *Universidad Complutense en El Escorial*, en los de *la Universidad Nacional de Educación a Distancia en Ávila y* en los de la *Fundación Universidad de Verano de Castilla y León en Segovia*. En general, esta colección pretende dar a conocer todas aquellas actividades que desarrollamos durante el período estival.

Se divide en cuatro series, dedicadas las tres primeras a la Educación Secundaria (la tercera a F.P.), y la cuarta a Infantil y Primaria.

Colección **Aulas de Verano**, que se identifica con el color "bermellón Salamanca"

- Serie "Ciencias" Color verde
- Serie "Humanidades" Color azul
- Serie "Técnicas" Color naranja
- Serie "Principios" Color amarillo

La segunda colección se denomina **Conocimiento Educativo.** Con ella pretendemos tanto difundir investigaciones realizadas por el profesorado o grupos de profesores, como dar a conocer aquellas acciones educativas que desarrolla el Instituto Superior de Formación del Profesorado durante del año académico.

La primera serie está dedicada fundamentalmente a investigación didáctica y, en particular, a las didácticas específicas de cada disciplina; la segunda serie se dirige al análisis de la situación educativa y estudios generales, siendo esta serie el lugar donde se darán a conocer nuestros Congresos EN_CLAVE DE CALID@D; la tercera serie, "Aula Permanente", da a conocer los distintos cursos que realizamos durante el período docente, y la cuarta serie, como su nombre indica, se dedica a estudios, siempre desde la perspectiva de la educación, sobre nuestro Patrimonio.

Colección **Conocimiento Educativo**, que se identifica con el color "amarillo oficial"

- Serie "Didáctica" Color azul
  Dentro de esta serie se publican los cinco anuarios europeos "Eulde", revistas de alta investigación en Didáctica de las Matemáticas, de las Lenguas, de las Ciencias Experimentales, de la Historia, la Geografía y las Ciencias Sociales y de las Expresiones (Plástica, Musical y Corporal) Se publican simultáneamente en castellano, francés, italiano, portugués e inglés.
- Serie "Situación" Color verde
- Serie "Aula Permanente" Color rojo
- Serie "Patrimonio" Color violeta

Estas colecciones, como hemos señalado, tienen un carácter de difusión y extensión educativa, que prestará un servicio a la intercomunicación, como hemos dicho también, entre los docentes que desarrollan sus tareas en las distintas Comunidades y Ciudades Autónomas de nuestro Estado. Pero, también, se pretende con ellas establecer un vehículo del

máximo rigor científico y académico en el que encuentren su lugar el trabajo, el estudio, la reflexión y la investigación de todo el profesorado español, de todos los niveles, sobre la problemática educativa.

Esta segunda función es singularmente importante, porque incentiva en los docentes el imprescindible objetivo investigador sobre la propia función, lo que constituye la única vía científica y, por tanto, con garantías de eficacia, para el más positivo desarrollo de la formación personal y los aprendizajes de calidad en los niños y los jóvenes españoles.

Índices de calidad de las publicaciones:

Todos los proyectos de publicación, en cualquiera de las dos colecciones, estarán avalados por cinco informes razonados, emitidos cada uno por un Profesor Doctor de reconocido prestigio de diferente centro, docente o de investigación, español o del extranjero. Al menos tres de los cinco informantes han de ser Catedráticos de Universidad, y al menos tres de los cinco centros han de ser españoles.

Los programas de publicación son aprobados por una comisión compuesta por el Director del Instituto Superior de Formación del Profesorado, la Directora de Programas y la Directora de Publicaciones del Instituto Superior de Formación del Profesorado y los Directores (o persona en quien deleguen) del Servicio de Publicaciones del Ministerio de Educación, Cultura y Deporte y del INCE.

NORMAS DE EDICIÓN DEL INSTITUTO SUPERIOR DE FORMACIÓN DEL PROFESORADO:

- Los artículos han de ser inéditos.
- Se entregarán en papel y se añadirá una copia en disquete (en un procesador de textos tipo Word).
- El autor/es debe dar los datos personales siguientes: referencia profesional, dirección y teléfono personal y del trabajo. En caso de trabajos colectivos, se referenciarán estos datos de todos los autores.
- Debe haber, al principio de cada artículo, un recuadro con un índice de los temas que trata el mismo.
- El autor debe huir de textos corridos y utilizar con la frecuencia adecuada, epígrafes y subepígrafes que aparezcan distribuidos en el texto, al menos, en cada doble página.
- Cuando se reproduzcan textos de autores, se entrecomillarán y se pondrán en cursiva.
- Al citar un libro, siempre debe aparecer la página de la que se toma la cita, excepto si se trata de un comentario general.
- Se deben adjuntar fotografías, esquemas, trabajos de alumnos,... que ilustren o expliquen el contenido del texto.
- Se debe adjuntar en un listado numerado correlativamente, las notas que se van a poner a pie de página, según las referencias incluidas en el texto.
- Al final de cada artículo, se adjuntará la lista de la bibliografía utilizada.
- La bibliografía debe ser citada de la siguiente manera: apellidos/s (con mayúsculas), coma; nombre según aparezca en el libro(en letra corriente), punto; título del libro en cursiva, punto; editorial, punto; ciudad de edición, coma y fecha de publicación, punto. Así se realizarán también las citas a pie de página.

CENTRAL DE EDICIONES DEL INSTITUTO SUPERIOR DE FORMACIÓN DEL PROFESORADO

- **Dirección y coordinación (I.S.F.P.):**
  Paseo del Prado 28, 6ª planta. 28014. Madrid. Teléfono: 91.506.57.17.

- **Suscripciones y distribución:**
  Instituto de Técnicas Educativas. C/ Alalpardo s/n. 28806. Alcalá de Henares. Teléfono: 91.889.18.50.

- **Puntos de venta:**
  - Ministerio de Educación, Cultura y Deporte. C/Alcalá, 36. Madrid.
  - Subdirección General de Información y Publicaciones del Ministerio de Educación, Cultura y Deporte. C/Juán del Rosal s/n. Madrid.

TÍTULOS EDITADOS

	COLECCIÓN	SERIE
La Educación Artística, clave para el desarrollo de la creatividad	AULAS DE VERANO	Principios
La experimentación en la enseñanza de las Ciencias	AULAS DE VERANO	Principios
Metodología en la enseñanza del Inglés	AULAS DE VERANO	Principios
Destrezas comunicativas en la Lengua Española	AULAS DE VERANO	Principios
Dificultades del aprendizaje de las Matemáticas	AULAS DE VERANO	Principios
La Geografía y la Historia, elementos del Medio	AULAS DE VERANO	Principios
La enseñanza de las Matemáticas a debate: referentes europeos	AULAS DE VERANO	Ciencias
El lenguaje de las Matemáticas en sus aplicaciones	AULAS DE VERANO	Ciencias
La iconografía en la enseñanza de la Historia del Arte	AULAS DE VERANO	Humanidades
Grandes avances de la Ciencia y la Tecnología	AULAS DE VERANO	Técnicas
EN_CLAVE DE CALID@D: la Dirección Escolar	CONOCIMIENTO EDUCATIVO	Situación
Felipe V y el Palacio Real de La Granja	CONOCIMIENTO EDUCATIVO	Patrimonio
Didáctica de la poesía en la Educación Secundaria	CONOCIMIENTO EDUCATIVO	Didáctica
La seducción de la lectura en edades tempranas	AULAS DE VERANO	Principios
Aplicaciones de las nuevas tecnologías en el aprendizaje de la Lengua Castellana	AULAS DE VERANO	Principios

Lenguas para abrir camino	AULAS DE VERANO	Principios
La dimensión artística y social de la ciudad	AULAS DE VERANO	Humanidades
La Lengua, vehículo cultural multidisciplinar	AULAS DE VERANO	Humanidades
Lenguas extranjeras: hacia un nuevo marco de referencia en su aprendizaje	AULAS DE VERANO	Humanidades
Globalización, crisis ambiental y Educación	AULAS DE VERANO	Ciencias
Los fundamentos teórico-didácticos de la Educación Física	CONOCIMIENTO EDUCATIVO	Didáctica
Los lenguajes de la expresión	AULAS DE VERANO	Principios
La comunicación literaria en las primeras edades	AULAS DE VERANO	Principios
La Física y la Química: del descubrimiento a la intervención	AULAS DE VERANO	Ciencias
La estadística y la probabilidad en el Bachillerato	CONOCIMIENTO EDUCATIVO	Didáctica
La estadística y la probabilidad en la Educación Secundaria Obligatoria	CONOCIMIENTO EDUCATIVO	Didáctica
Nuevas profesiones para el servicio a la sociedad	AULAS DE VERANO	Técnicas

9 788436 936605

La primera de nuestras colecciones se denomina **Aulas de Verano**, y pretende que todo el profesorado pueda acceder al conocimiento de las conferencias, ponencias, mesas redondas, talleres y actividades profesionales docentes que se desarrollan durante los veranos en la *Universidad Internacional Menéndez Pelayo de Santander*, en los cursos de la *Universidad Complutense en El Escorial*, en los de la *Universidad Nacional de Educación a Distancia en Ávila y* en los de la *Fundación de Universidades de Castilla y León en Segovia*.

Colección **Aulas de Verano**, que se identifica con el color «bermellón Salamanca»	
• Serie «Ciencias»	Color verde
• Serie «Humanidades»	Color azul
• Serie «Técnicas»	Color naranja
• Serie «Principios»	Color amarillo

La segunda colección se denomina **Conocimiento Educativo.** Con ella pretendemos tanto difundir investigaciones realizadas por el profesorado o grupos de profesores, como dar a conocer aquellas acciones educativas que desarrolla el Instituto Superior de Formación del Profesorado durante el año académico.

Colección **Conocimiento Educativo**, que se identifica con el «color amarillo oficial»	
• Serie «Didáctica»	Color azul
Dentro de esta serie se publican los cinco anuarios europeos "Eulde", revistas de alta investigación en Didáctica de las Matemáticas, de las Lenguas, de las Ciencias Experimentales, de la Historia, la Geografía y las Ciencias Sociales y de las Expresiones (Plástica, Musical y Corporal) Se publican simultáneamente en castellano, francés, italiano, portugués e inglés.	
• Serie «Situación»	Color verde
• Serie «Aula Permanente»	Color rojo
• Serie «Patrimonio»	Color violeta

Estas colecciones tienen un carácter de difusión y extensión educativa, al servicio de la intercomunicación entre los docentes que desarrollan sus tareas en las distintas Comunidades y Ciudades Autónomas de nuestro Estado.

MINISTERIO
DE EDUCACIÓN,
CULTURA Y DEPORTE