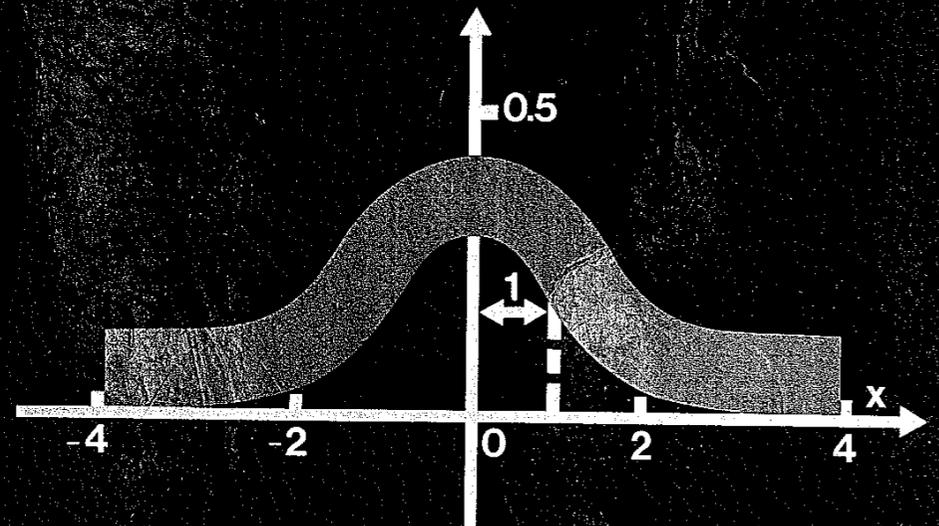


J.M. Bernardo

BIOESTADISTICA

una perspectiva Bayesiana



20
138
246

BIOESTADISTICA
Una Perspectiva Bayesiana

Dirección de edición: Albert Vicens

A mis primeros maestros,

LUCRECIA HERRANZ
y GREGORIO BERNARDO

Primera edición. 1981

Depósito Legal: B. 35.455-1981

ISBN: 84-316-1889-2

N.º de Orden V.V.: B-928

© J.M. BERNARDO
Sobre la parte literaria

Reservados todos los derechos de edición a favor de Ediciones Vicens-Vives, S.A.
Prohibida la reproducción total o parcial por cualquier medio.

IMPRESO EN ESPAÑA
PRINTED IN SPAIN

Editado por Ediciones VICENS-VIVES, S.A. Avda. de Sarriá, 130. Barcelona-17.
Impreso por Gráficas INSTAR S.A. Constitución, 19. Barcelona-14.



El autor

Es Director del Departamento de Bioestadística de la Universidad de Valencia, miembro electo del European Committee of Statisticians y miembro electo del International Statistical Institute

La densa trayectoria académica del Profesor Bernardo se inicia en 1976 con una tesis doctoral por la Universidad de Londres galardonada con el segundo premio internacional de tesis doctorales de la especialidad y se consolida con una Fellowship postdoctoral en la Universidad de Yale.

Creador de un centro de investigación en metodología Bayesiana internacionalmente conocido, organizó en Valencia, en 1979, el primer congreso mundial sobre métodos estadísticos Bayesianos.

Sus conferencias en Universidades extranjeras (Harvard, Yale, New York, Chicago, Vancouver, México, Londres, Roma, Florencia, Bolonia, Frankfurt, etc.), comunicaciones en congresos internacionales (Luxemburgo, Praga, Varsovia, Grenoble, Lovaina, Valencia, Roma, Varna, Brighton, etc.) y sus numerosas publicaciones en revistas especializadas del máximo nivel internacional (*Annals of Statistics*, *Journal of the Royal Statistical Society*, etc.), han dado prestigio a la Estadística Española en el mundo.

Prefacio

La transición de los métodos estadísticos clásicos a los métodos Bayesianos en la práctica profesional y en la investigación no está siendo tan rápida como algunos de nosotros hubiéramos deseado. Una de las causas de este retraso es posiblemente la falta de libros de texto elementales que muestren a los profesionales, con ejemplos escogidos de su propio campo, la forma de utilizar la metodología Bayesiana en su trabajo. Este libro pretende contribuir a llenar esa laguna.

La metodología Bayesiana resulta atractiva por varios motivos. En primer lugar, y a diferencia de la metodología clásica, está exenta de contradicciones internas debido a su planteamiento axiomático. En segundo lugar, proporciona una unidad conceptual y una metodología definida que elimina la necesidad de procedimientos *ad hoc*. Finalmente, la metodología Bayesiana puede utilizarse para precisar el contenido y perfilar el alcance de los métodos clásicos; en efecto, muchos de los resultados clásicos más conocidos admiten una interpretación Bayesiana, pero otros muchos violan los principios básicos de coherencia en que los métodos Bayesianos se fundamentan y resultan por lo tanto inadmisibles para quienes encuentren razonables tales principios.

Las técnicas Bayesianas son además especialmente relevantes para la práctica y para la investigación médica. En primer lugar, la mayor parte de los problemas con los que el médico debe enfrentarse son problemas de decisión y la metodología Bayesiana es la única que describe el proceso lógico que debe seguirse para decidir, de forma razonable, sobre la mejor forma de actuar. En segundo lugar, los médicos cuentan típicamente con una importante cantidad de información inicial, y las técnicas Bayesianas son las únicas que permiten utilizarla. Finalmente, los conceptos básicos de la metodología Bayesiana son mucho más intuitivos que sus equivalentes clásicos y su correcta asimilación resulta posible sin recurrir a matemáticas sofisticadas.

Este libro es una introducción a los métodos estadísticos Bayesianos, que subraya su fundamentación axiomática y el tratamiento unificado de los conceptos de probabilidad, inferencia, y decisión a que da lugar. El énfasis se sitúa en los conceptos, de forma que aunque se discuten muchos problemas específicos, el objetivo primordial es dotar al lector de una visión de conjunto. El análisis de datos experimentales en situaciones concretas y la descripción de las aplicaciones más comunes de los métodos estadísticos en Medicina y Biología serán objeto de tratamiento posterior.

El contenido de este libro es, desde mi punto de vista, lo mínimo que un profesional debería saber sobre el proceso lógico que debe presidir el análisis de la información disponible y la toma de decisiones a que dicho análisis pueda dar lugar. Los únicos requisitos previos son aquellos conocimientos elementales de matemáticas, que suelen haber sido adquiridos en la enseñanza secundaria; se trata de un curso de introducción que no requiere conocimientos previos de estadística ni exige recurrir a otras fuentes.

El libro está dividido en siete capítulos. Ecuaciones, definiciones, teoremas y ejemplos están numerados en notación decimal; así, por ejemplo, ecuación 2.3.1 se refiere a la ecuación (1) de la Sección 3 del Capítulo 2. Entre los apéndices finales, se incluyen índices numéricos de ecuaciones, definiciones, teoremas y ejemplos que facilitan su localización.

Cada capítulo empieza con un corto resumen de su contenido y termina con una sección bibliográfica y una lista de problemas propuestos cuyos resultados numéricos aparecen en un apéndice. La solución de algunos de estos problemas requiere el uso de unas tablas estadísticas; las de Ferrándiz (1980) están especialmente orientadas hacia la metodología Bayesiana.

Son muchos los autores responsables de los conceptos y resultados expuestos en este libro; a lo largo del texto se hace referencia específica a sus trabajos. Entre el material expuesto, el lector familiarizado con la literatura especializada reconocerá varias contribuciones originales, generalmente relacionadas con las medidas de información y con su uso en la determinación de distribuciones de referencia y en el diseño de experimentos.

Numerosas personas han contribuido directamente a hacer posible este libro. En mi formación profesional, ha sido determinante la influencia del Profesor Dennis V. Lindley, bajo cuya dirección realicé la tesis doctoral en el *University College* de Londres. El estímulo inicial para escribir el libro vino de Albert Vicens y de mis alumnos de la Facultad de Medicina de Valencia a quienes les he impartido este curso durante tres años consecutivos. Javier Girón y Miguel-Ángel Gómez-Villegas han leído el último manuscrito y debo agradecerles sus útiles comentarios y la corrección de algunos errores. La contribución de mis compañeros de trabajo ha sido muy importante: José Bermudez, Juan Ferrándiz, Maite Rabena, Luis Sanjuan y Mario Sendra

han leído los sucesivos manuscritos, han preparado gráficas y tablas, han resuelto los problemas, y han aportado valiosas sugerencias; Charo Escandón ha mecanografiado cuidadosamente las sucesivas versiones del texto. A todos ellos les agradezco su esfuerzo. Todos ellos comparten los méritos que pueda tener este libro, aunque sólo yo soy responsable de los defectos que permanezcan en él.

Valencia, abril 1980

JOSÉ-MIGUEL BERNARDO

Índice

Prefacio	IX
Capítulo 1	
Introducción	1
1.1 Alcance y objetivos del libro	1
1.2 Estadística y Teoría de la Decisión	3
1.3 Probabilidad	4
1.4 Inferencia estadística	5
1.5 Problemas específicos de decisión	6
1.6 Estructura del libro	7
Capítulo 2	
Fundamentos de la Estadística y de la Teoría de la Decisión	9
2.1 Estructura de un problema de decisión	10
Ejemplo 2.1.1 Oportunidad de una operación	13
Ejemplo 2.1.2 Elección de un medio de transporte	14
2.2 Solución intuitiva a un problema de decisión	16
Ejemplo 2.2.1 Participación en una lotería	20
Ejemplo 2.2.2 Elección de un medio de transporte (cont.)	21
2.3 Principios de coherencia	22
Ejemplo 2.3.1 Opciones económicas alternativas	27
2.4 Probabilidad como grado de creencia	28
Ejemplo 2.4.1 Temperaturas	32
2.5 Maximización de la utilidad esperada	32
Ejemplo 2.5.1 Opciones económicas alternativas (cont.)	35

2.6	Otros criterios de decisión	36
	Ejemplo 2.6.1 Forma de estudio	40
2.7	Discusión y referencias	41
	Problemas	43
Capítulo 3		
Medida de probabilidad		
3.1	Espacios probabilísticos	45
	Ejemplo 3.1.1 Número de hemáties en la sangre	47
3.2	Teoremas básicos	48
	Ejemplo 3.2.1 Problema de cumpleaños	51
	Ejemplo 3.2.2 Equipos de guardia	53
3.3	Independencia e intercambiabilidad	53
	Ejemplo 3.3.1 Ruleta	54
	Ejemplo 3.3.2 Sexo de un recién nacido	56
	Ejemplo 3.3.3 Aparición de tumores	57
3.4	Teoremas de Bayes y de la probabilidad total	58
	Ejemplo 3.4.1 Premio literario	59
	Ejemplo 3.4.2 Test de tuberculina	61
	Ejemplo 3.4.3 Diagnósis	62
3.5	Análisis secuencial	63
	Ejemplo 3.5.1 Prueba del alcohol	63
	Ejemplo 3.5.2 Valor de la información	66
3.6	Asignación de probabilidades	69
	Ejemplo 3.6.1 Pronóstico de un partido	70
	Ejemplo 3.6.2 Calificación de exámenes	72
3.7	Discusión y referencias	73
	Problemas	76
Capítulo 4		
Cantidades aleatorias		
4.1	Cantidad aleatoria y función de distribución	80
	Ejemplo 4.1.1 Temperaturas	81
	Ejemplo 4.1.2 Temperaturas (cont.)	82
4.2	Distribuciones discretas	83
	Ejemplo 4.2.1 Elección de pacientes	84
	Ejemplo 4.2.2 Sexo de un recién nacido (cont.)	85
	Ejemplo 4.2.3 Probabilidad de cáncer	87
4.3	Distribuciones continuas	88
	Ejemplo 4.3.1 Longitud de una úlcera	92
	Ejemplo 4.3.2 Pesos	95
4.4	Funciones de una cantidad aleatoria	97
	Ejemplo 4.4.1 Pacientes hospitalizados	97
	Ejemplo 4.4.2 Normalización de una distribución Beta	99

	Ejemplo 4.4.3 Test de normalidad	100
	Ejemplo 4.4.4 Simulación de observaciones normales	102
4.5	Características de una distribución	102
	Ejemplo 4.5.1 Elección de pacientes (cont.)	105
	Ejemplo 4.5.2 Tiempos de espera	106
	Ejemplo 4.5.3 Normalización de una distribución Beta (cont.)	109
	Ejemplo 4.5.4 Tiempos de espera (cont.)	111
4.6	Distribuciones multivariantes	112
	Ejemplo 4.6.1 Tipos de enfermedad	114
	Ejemplo 4.6.2 Tipos de enfermedad (cont.)	115
	Ejemplo 4.6.3 Distribución del cociente	117
	Ejemplo 4.6.4 Pesos y estaturas	122
4.7	Discusión y referencias	123
	Problemas	124
Capítulo 5		
El proceso de aprendizaje		
5.1	Cuantificación de la información inicial	128
	Ejemplo 5.1.1 Diagnósis	128
	Ejemplo 5.1.2 Cantidad de tirosina	129
	Ejemplo 5.1.3 Cantidad de tirosina (cont.)	130
	Ejemplo 5.1.4 Tasas de morbilidad	132
5.2	Función de verosimilitud	134
	Ejemplo 5.2.1 Diagnósis (cont.)	134
5.3	Distribución predictiva	139
	Ejemplo 5.3.1 Diagnósis (cont.)	140
	Ejemplo 5.3.2 Tasas de morbilidad (cont.)	142
	Ejemplo 5.3.3 Cantidad de tirosina (cont.)	144
	Ejemplo 5.3.4 Cantidad de tirosina (cont.)	144
5.4	Teorema de Bayes y distribución final	145
	Ejemplo 5.4.1 Diagnósis (cont.)	148
	Ejemplo 5.4.2 Tasa de morbilidad (cont.)	150
	Ejemplo 5.4.3 Tasa de morbilidad (cont.)	151
	Ejemplo 5.4.4 Cantidad de tirosina (cont.)	154
5.5	Parámetros marginales	155
	Ejemplo 5.5.1 Diagnósis	156
	Ejemplo 5.5.2 pH de la saliva	158
5.6	Predicción	159
	Ejemplo 5.6.1 Diagnósis (cont.)	160
	Ejemplo 5.6.2 Tasas de morbilidad (cont.)	161
	Ejemplo 5.6.3 Cantidad de tirosina (cont.)	163
5.7	Discusión y referencias	163
	Problemas	164

Capítulo 6

Métodos aproximados de inferencia	167
6.1 Descripción de la distribución final	168
Ejemplo 6.1.1 Tasas de morbilidad (cont.)	168
Ejemplo 6.1.2 Cantidad de tirosina (cont.)	172
6.2 Familias conjugadas de distribuciones	173
Ejemplo 6.2.1 Tiempos de espera	177
6.3 Aproximación normal a la distribución final	178
Ejemplo 6.3.1 Aproximación Poisson de una Binomial	179
Ejemplo 6.3.2 Ajuste de distribuciones gamma	182
Ejemplo 6.3.3 Cálculo de probabilidades en distribuciones gamma	182
6.4 Comportamiento asintótico de la distribución final	183
Ejemplo 6.4.1 Probabilidad de contagio	186
Ejemplo 6.4.2 Composición del líquido cefalorraquídeo	189
6.5 Distribuciones finales de referencia	190
Ejemplo 6.5.1 Diagnóstico	191
Ejemplo 6.5.2 Incidencia de una enfermedad rara	194
Ejemplo 6.5.3 Composición de la orina	196
6.6 Análisis de sensibilidad	198
Ejemplo 6.6.1 Resultados de un muestreo	200
6.7 Discusión y referencias	202
Problemas	204

Capítulo 7

Análisis cuantitativo de decisiones	207
7.1 Contraste de hipótesis y estimación puntual	207
Ejemplo 7.1.1 Comercialización de un fármaco	209
Ejemplo 7.1.2 Control de calidad	212
7.2 La inferencia estadística como problema de decisión	212
7.3 Evaluación de utilidades	215
Ejemplo 7.3.1 Utilidades monetarias	219
7.4 Valor esperado de la información	219
Ejemplo 7.4.1 Valor diagnóstico de un test	221
7.5 Diseño de experimentos	224
Ejemplo 7.5.1 Exploraciones peligrosas	227
Ejemplo 7.5.2 Tamaño óptimo de una encueseta	230
7.6 Algunos problemas médicos de decisión	232
7.6.1 Calibrado	232
7.6.2 Diagnóstico	234
7.6.3 Elección de tratamiento	235
7.7 Discusión y referencias	236
Problemas	238

Referencias

Soluciones a los problemas	241
Índice de conceptos	247
Índice de autores	253
Índice de teoremas	257
Índice de ecuaciones	259
Índice de definiciones	261
Índice de ejemplos	263
Índice de símbolos	265
	267

Introducción

Los métodos estadísticos, en su sentido más amplio, constituyen una herramienta de trabajo cada vez más importante en los campos profesionales más diversos. Difícilmente puede ser considerada completa hoy la formación de biólogos, economistas, físicos, ingenieros, militares, políticos, psicólogos, químicos, o cualquier otro grupo profesional, sin la asimilación de unos conceptos básicos de metodología estadística que les permitan analizar adecuadamente los datos sobre los que trabajan y tomar de forma racional las decisiones oportunas.

Aunque los problemas específicos pueden ser muy distintos, existe una metodología estadística general que permite abordar de forma sistemática cualquier problema concreto de decisión o de análisis de datos.

Nuestro propósito es proporcionar al lector una introducción a la metodología estadística moderna concebida como un instrumento para *analizar* adecuadamente la información de que se dispone y *decidir*, de manera razonable, sobre la mejor forma de actuar.

1.1. Alcance y objetivos del libro

La importancia práctica de saber analizar la información de que se dispone y de saber utilizarla para tomar decisiones de forma razonable, resulta indiscutible cuando se observa la continua cadena de decisiones sobre la que descansa toda actividad humana. Este libro está dirigido a cualquier universitario que desee una formación moderna en metodología estadística y teoría de la decisión.

En particular, un elemento característico de la vida profesional de un médico es la constante toma de decisiones en ambiente de incertidumbre; decisiones, por ejemplo, sobre el diagnóstico más verosímil, la oportunidad de una intervención quirúrgica, el tratamiento más adecuado o la eficacia de un programa de inmunización. De hecho, muchos de los ejemplos contenidos en este libro, con los que pretendemos facilitar la asimilación de los conceptos expuestos, han sido extraídos de la experiencia del autor en las aplicaciones a este campo, por lo que el libro resulta especialmente indicado para médicos y estudiantes de Medicina.

El lector que haya asimilado las ideas que vamos a exponer estará en condiciones de

- (i) *Extraer las conclusiones pertinentes de un conjunto de datos con estructura sencilla*
- (ii) *Resolver problemas de decisión moderadamente complicados*
- (iii) *Evaluar el contenido de las conclusiones de tipo estadístico publicadas en la literatura científica.*

Para lograr estos objetivos, no es suficiente la lectura del texto; su estudio debe ser completado con la realización de los problemas propuestos al final de cada capítulo. Para facilitar esta tarea, las soluciones numéricas a estos problemas han sido incluidas en un apéndice final. Todos los ejemplos y problemas contenidos en este libro pueden resolverse con unas tablas estadísticas adecuadas y una simple calculadora de bolsillo que contenga las funciones elementales. Sin embargo, resulta cómodo disponer de un modelo de calculadora que permita el cálculo directo de medias y desviaciones típicas, de la función *factorial* (ver Sección 3.3), de la *función de distribución normal* (ver Sección 4.3) y de su función inversa. Numerosas marcas de calculadoras ofrecen modelos, diseñados para aplicaciones estadísticas, que disponen de estas funciones.

Frecuentemente, el lector se encontrará con la necesidad de repasar conceptos que ya creía asimilados; los resúmenes situados al principio de cada capítulo y los índices alfabéticos de conceptos y de notación situados al final del libro pueden ayudarle a hacerlo. Las *Tablas Estadísticas* de Ferrándiz (1980) contienen además un útil formulario que recoge la mayor parte de los resultados expuestos en este libro.

La correcta asimilación de los conceptos expuestos en este libro exige del lector la preparación matemática que en España debe ser obtenida en BUP y COU. Específicamente, resulta necesario conocer algunas funciones elementales, como el logaritmo, la función exponencial o las funciones trigonométricas; saber resolver ecuaciones sencillas y sistemas lineales; conocer el álgebra matricial, y tener nociones de cálculo diferencial e integral. Si el

lector no se siente seguro de sus conocimientos en alguna de estas áreas, puede actualizarlos con ayuda del excelente texto de García-García y López-Pellicer (1977). No obstante, los párrafos con mayor dificultad matemática han sido impresos en letra más pequeña para su fácil identificación, y el texto ha sido redactado de forma que estos párrafos pueden ser omitidos en primera lectura sin pérdida de continuidad.

En este libro se desarrollan los conceptos fundamentales de la metodología estadística. Aunque, naturalmente, se obtienen las fórmulas específicas que corresponden a las aplicaciones más frecuentes, el énfasis se sitúa en la exposición de una metodología general, capaz de permitir al lector abordar por sí mismo el análisis de cualquier problema.

Al final de cada capítulo, se discute la relevancia y el alcance de las ideas introducidas en él y se dan referencias para aquellos lectores interesados en aplicaciones más concretas o más detalladas, desarrollos teóricos más completos, generalizaciones de los métodos expuestos o estudios de su desarrollo histórico. La lista completa de tales referencias aparece al final del libro. Aunque ciertamente este conjunto de referencias no es exhaustivo, es lo suficientemente extenso como para ofrecer una amplia panorámica de casi todos los aspectos de la estadística y de la teoría de la decisión.

1.2. Estadística y Teoría de la Decisión

La metodología estadística clásica aparece ante el investigador como un conjunto de *recetas* cada una de las cuales resulta apropiada en determinados casos y bajo ciertas condiciones. Sin embargo, hoy ya resulta posible ofrecer, incluso a un nivel elemental, una metodología unificada, totalmente general, que se deriva de analizar cuidadosamente el proceso lógico que debe seguirse para tomar una decisión.

En efecto, la teoría de la decisión no sólo proporciona una metodología para resolver problemas de decisión concretos sino que proporciona además una base teórica sobre la que puede construirse una metodología general unificada que contiene, como casos particulares, aquellas recetas clásicas cuya utilidad ha sido demostrada por el tiempo.

Empezaremos por situarnos ante el problema general de decisión, es decir ante el problema de la elección razonada entre un determinado conjunto de alternativas, en presencia de incertidumbre sobre algunos de los factores que condicionan las consecuencias de tal elección. Nos preguntaremos cuál es el proceso lógico que debe seguirse para tomar una decisión. Para ello, no trataremos de describir la forma en que comúnmente se toman las decisiones. En su lugar, trataremos de encontrar unos *principios básicos* sobre los que edificar una teoría de la decisión. No nos interesa como se toman las decisiones

sino como se *deberían* tomar. No buscamos una teoría descriptiva sino una teoría normativa.

La conclusión fundamental de esta investigación es que existe una *única* forma razonable de tomar decisiones. Aunque naturalmente existe una cierta libertad en la elección de los principios básicos, el resultado es siempre el mismo. En primer lugar, es necesario determinar el conjunto de las decisiones posibles y el de aquellos sucesos cuya ocurrencia pueda modificar las consecuencias de la decisión tomada. En segundo lugar debe cuantificarse, mediante probabilidades, la verosimilitud asociada por el decisor a la ocurrencia de tales sucesos. En tercer lugar deben describirse detalladamente las posibles consecuencias de cada una de las decisiones, y las preferencias del decisor entre ellas deben ser evaluadas y cuantificadas en términos de una magnitud común que recibe el nombre de utilidad. Finalmente, debe tomarse aquella decisión que, con base a las probabilidades calculadas, proporcione la máxima utilidad esperada. La fuerza del *debe* utilizado varias veces en este párrafo radica en que cualquier desviación de tales preceptos está en contradicción con los principios básicos de partida. En consecuencia, cualquier otro criterio de decisión resulta inadmisibles para quien acepte tales principios.

1.3. Probabilidad

La mayor parte de los libros de estadística dan la impresión de que no existe controversia alguna sobre la validez de la metodología estadística en uso. Esto es simplemente falso. En los años cincuenta empezó a consolidarse la idea, expuesta en la sección anterior, de que la metodología estadística debería tener unos fundamentos axiomáticos sólidos sobre los que apoyarse, fundamentos que fueron encontrados en la teoría de la decisión. Una consecuencia importante de tales fundamentos es que resulta necesario cuantificar nuestra información sobre la verosimilitud de los distintos sucesos inciertos que resulten relevantes en cada problema concreto, y que tal cuantificación debe ser realizada necesariamente mediante una medida que satisfaga determinadas propiedades, a la que llamamos *probabilidad*. La idea de que la probabilidad de un suceso en unas condiciones determinadas no es más que una medida de la verosimilitud asociada por el decisor a la ocurrencia de tal suceso en esas condiciones, caracteriza la escuela de pensamiento en que este libro se inscribe, y que por motivos que luego mencionaremos recibe el nombre de *Bayesiana*. Los métodos estadísticos *clásicos*, por otra parte, limitan el concepto de probabilidad a aquellos sucesos en los que pueden definirse frecuencias relativas, lo que reduce seriamente su utilidad práctica, y no disponen de una base axiomática en la que fundamentarse, por lo que frecuentemente dan lugar a resultados contradictorios.

Las consecuencias matemáticas de la definición de probabilidad resultan,

no obstante, independientes de la interpretación específica que se dé a este concepto de forma que, aunque existen varias interpretaciones del *concepto* de probabilidad, puede hablarse de una única *Teoría de la Probabilidad*. Estudiaremos, pues, la forma de asignar probabilidades, y analizaremos aquellas propiedades matemáticas de la medida de probabilidad que puedan facilitar el cálculo de unas probabilidades a partir de los valores de otras.

1.4. Inferencia estadística

La reacción natural de cualquiera que deba tomar una decisión en presencia de incertidumbre es eliminar cuanta incertidumbre le sea posible obteniendo más información. En efecto, frente a un problema de decisión específico, el decisor empieza por precisar la información de que inicialmente dispone sobre las incertidumbres que afectan a la situación; procura entonces obtener nuevos datos que puedan proporcionarle información relevante y, finalmente, combina esta nueva información con aquella de que inicialmente disponía para tomar entonces la decisión más apropiada.

Más concretamente, la información inicial, descrita por *probabilidades iniciales*, es combinada con los datos mediante el llamado *Teorema de Bayes* para obtener las *probabilidades finales*, que describen la información total de que se dispone. Este proceso, que denominaremos *proceso de aprendizaje*, constituye la esencia de la *metodología Bayesiana*, cuyo nombre hace referencia precisamente al uso repetido del teorema de Bayes.

A la mayor parte de los lectores les resultará familiar la expresión plural *estadísticas* con la que suele designarse a determinados conjunto ordenados de datos, obtenidos por diversos procedimientos. Se conoce como *Estadística Descriptiva* al conjunto de técnicas que facilitan la organización, resumen y comunicación de estos datos. Las más elementales de estas técnicas, que probablemente conocerá el lector, son la obtención de medidas de localización y de dispersión como la media o la desviación típica, y la construcción de determinadas representaciones gráficas, como los histogramas. El lector interesado en releer estos conceptos puede recurrir, por ejemplo, a los primeros capítulos del texto de Spiegel (1961). En esta línea, pero a un nivel mucho más sofisticado, se encuentra el libro de Tukey (1977).

En este volumen apenas nos ocuparemos de estas técnicas descriptivas. En su lugar, utilizando la metodología mencionada al principio de esta sección, abordaremos la construcción de una teoría general que nos permita precisar las *conclusiones* que pueden ser extraídas de los datos. *Inferencia Estadística* es el nombre con que generalmente se conoce al conjunto de métodos que se proponen extraer conclusiones que van más allá de la mera descripción de los datos.

Cuando *no* se dispone de información inicial, los métodos clásicos y los métodos Bayesianos de inferencia llegan frecuentemente a las mismas conclusiones numéricas, aunque con interpretaciones muy diferentes. Esta coincidencia numérica permitirá al lector de este libro dar un sentido razonable a muchas de las *recetas* tradicionales. Cuando, por el contrario, y este es el caso más frecuente, se dispone de información inicial, no existen realmente métodos alternativos; la metodología Bayesiana proporciona la *única* forma sistemática de incorporar esta información en el análisis.

1.5. Problemas específicos de decisión

Como ya hemos mencionado, una consecuencia de los principios básicos en que se fundamenta la teoría de la decisión es que, en cada problema concreto, el decisor debe evaluar mediante probabilidades la verosimilitud, *en el momento de tomar la decisión*, de los distintos sucesos inciertos presentes en el problema. Estas probabilidades deben reflejar *toda* la información de que se dispone en el momento de tomar la decisión. Frecuentemente, esta información estará compuesta por la información inicial de que disponía el decisor y por la proporcionada por un conjunto de datos adicionales; en este caso, las probabilidades que se requieren son las probabilidades *finales* mencionadas en la sección anterior. Consecuentemente, para resolver un problema de decisión es frecuentemente necesario resolver primero un problema de inferencia.

También hemos mencionado ya, que en cada problema de decisión concreto el decisor debe cuantificar sus preferencias entre las posibles consecuencias de sus decisiones. En general, no se trata de una tarea fácil, pero desarrollaremos varios procedimientos que facilitan su realización.

Frecuentemente, los problemas de decisión no se presentan aislados sino formando una cadena, de forma que una vez tomada una decisión se plantea un nuevo problema de decisión cuyas características dependen de las decisiones tomadas hasta entonces. Demostraremos, sin embargo, que el estudio de las *decisiones sucesivas* no plantea problemas nuevos; bastará resolver consecutivamente todos los problemas de decisión que integran la cadena, empezando por la decisión que deba ser tomada en *último* lugar. En particular, el problema del *diseño de experimentos* puede ser planteado como una cadena de dos decisiones consecutivas. Se trata, en efecto, de elegir el experimento más adecuado para, una vez realizado, tomar la decisión que resulte más apropiada a la vista de la información proporcionada por el experimento elegido.

1.6. Estructura del libro

El libro está dividido en siete capítulos, el primero de los cuales es esta Introducción.

En el capítulo segundo, *Fundamentos de la Inferencia y de la Teoría de la Decisión* se especifican los principios básicos de consistencia y se demuestra que la única forma de tomar decisiones compatibles con ellos es elegir la decisión que maximiza la utilidad esperada.

En el tercer capítulo, *Medida de Probabilidad*, se estudian las propiedades matemáticas de la medida que debe utilizarse para describir la verosimilitud de los sucesos inciertos relevantes y los métodos de que se dispone para la *asignación de probabilidades*. El estudio de la *Teoría de la Probabilidad* se completa en el capítulo cuarto, *Cantidades Aleatorias*, con el análisis de las probabilidades asociadas a los distintos valores numéricos que pueden tomar las magnitudes cuantificables.

En el capítulo quinto, *El Proceso de Aprendizaje*, se estudia la forma de incorporar al análisis cualquier información adicional y se analiza el tipo de conclusiones y predicciones que pueden ser extraídas de la información de que se dispone. El estudio de la *Inferencia Estadística* se completa en el capítulo sexto, *Métodos Aproximados de Inferencia*, con la descripción de las aproximaciones que pueden realizarse para facilitar el análisis y la comunicación de los resultados, y con el estudio de las conclusiones que pueden extraerse de los datos cuando no se dispone de información inicial.

En el séptimo y último capítulo, *Problemas Específicos de Decisión*, se estudian los problemas clásicos de *estimación puntual y contraste de hipótesis*, se aborda el problema de la *evaluación de las utilidades* y se discuten las cadenas de *decisiones sucesivas* y el *diseño de experimentos*. Finalmente, se comenta la estructura de algunos *problemas médicos de decisión*, concretamente los de calibrado, diagnóstico y elección de tratamiento.

La lista alfabética de los trabajos citados, las soluciones a los problemas propuestos, los índices numéricos de definiciones, ecuaciones, ejemplos y teoremas, y los índices alfabéticos de símbolos y de conceptos, completan este volumen.

Fundamentos de la estadística y de la teoría de la decisión

En este capítulo se estructura un problema de decisión y se discute el proceso lógico que debe seguirse para tomar una decisión. Esto proporciona el *fundamento* sobre el que descansan el concepto de probabilidad, los métodos de inferencia y los criterios de decisión y de diseño de experimentos que serán desarrollados en el resto del libro y que son globalmente conocidos como *métodos Bayesianos*.

El procedimiento seguido es *axiomático*. Se presentan y se defienden determinados principios de *comportamiento coherente* y se deduce de ellos la necesidad de asignar una *medida de probabilidad* que describa la información del decisor y una *función de utilidad* que describa sus preferencias, y la de elegir aquella decisión capaz de *maximizar la utilidad esperada*.

El capítulo concluye analizando críticamente los resultados obtenidos, comparándolos con los que se derivan de otros criterios de decisión propuestos en la literatura, y esbozando el desarrollo histórico de las ideas expuestas.

En 1926, F. P. Ramsey puso de manifiesto que unos pocos principios básicos sobre un comportamiento coherente en la elección entre opciones alternativas son suficientes para, tomados como axiomas, deducir de ellos una teoría general de la decisión y de la inferencia estadística. Como por otra parte, casi todos los problemas a cuya solución han contribuido tradicionalmente los métodos estadísticos, pueden reducirse a problemas de decisión en ambiente de incertidumbre, nuestra primera tarea será estudiar la estructura de tales problemas.

2.1. Estructura de un problema de decisión

Nos hallamos frente a un problema de decisión cuando debemos elegir entre dos o más formas de actuar. La mayor parte de nuestras decisiones son triviales, como cuando elegimos una película de la cartelera o el menú de un restaurante. Sin embargo, no es difícil imaginar problemas de decisión cuyas consecuencias son importantes y deben ser cuidadosamente consideradas antes de llegar a una conclusión; así, las decisiones relativas a un cambio de trabajo o a la compra de una casa exigen una seria reflexión.

Aunque sociólogos, historiadores y políticos han escrito a menudo sobre la forma en que determinadas decisiones se toman o han sido tomadas, se ha escrito muy poco sobre el tema de como *deberían* tomarse. La moderna teoría de la decisión propone un determinado método de tomar decisiones y demuestra además que es el único método compatible con unos pocos principios básicos sobre la *elección coherente* entre opciones alternativas.

Lo primero que hay que hacer cuando nos enfrentamos a un problema de decisión es considerar el conjunto de las posibles formas de actuación que se nos ofrecen. No es necesario distinguir entre una decisión y la acción a que da lugar. En efecto, si la acción no llega a realizarse es porque algo lo ha impedido dando lugar con ello a un nuevo problema de decisión. Generalmente, no resulta adecuado considerar únicamente una decisión y su negación como segunda decisión, formulando el problema con sólo dos alternativas. No es correcto, por ejemplo, plantearse si estudiar o no Medicina. En efecto, si uno decide no intentar ser médico tiene que hacer otra cosa para desarrollar su vida profesional; estudiar otra carrera, preparar unas oposiciones, buscar un trabajo. Existen en realidad muchas formas alternativas de desarrollarse profesionalmente y el problema de decisión consiste en una elección entre ellas y no en una simple comparación entre estudiar o no Medicina.

El primer paso para resolver el problema de decisión es, pues, elaborar el conjunto de las posibles alternativas, al que llamaremos *espacio de decisiones* y denotaremos por D . Debe ponerse especial atención en la construcción del espacio de decisiones porque el modelo que vamos a construir se limitará a elegir uno de sus elementos. Nunca puede uno estar totalmente seguro de que ha incluido en D todas las posibilidades interesantes; siempre puede aparecer un compañero ingenioso que nos señale una alternativa que no hemos considerado y nos obligue a replantear el problema. Un buen decisor debe tener la inventiva y el conocimiento del tema suficientes para elaborar un espacio de decisiones *exhaustivo*, es decir que agote todas las posibilidades que puedan, en principio, parecer razonables.

Es conveniente, asimismo, exigir que el espacio D de decisiones esté constituido por un conjunto de alternativas, de forma que la elección de uno de

los elementos de D *excluya* la elección de cualquier otro. Este requerimiento *no* supone pérdida de generalidad. Así, por ejemplo, para elegir los elementos opcionales que se desean en un nuevo coche, la lista de opciones ofrecidas por el fabricante no es un espacio de decisiones adecuado puesto que uno puede desear dos o más de las opciones ofrecidas, pero el conjunto de las partes de tal lista resulta serlo. De forma análoga, cualquier problema de decisión puede plantearse como el de la elección de un elemento, y uno sólo, de un conjunto aptopiado.

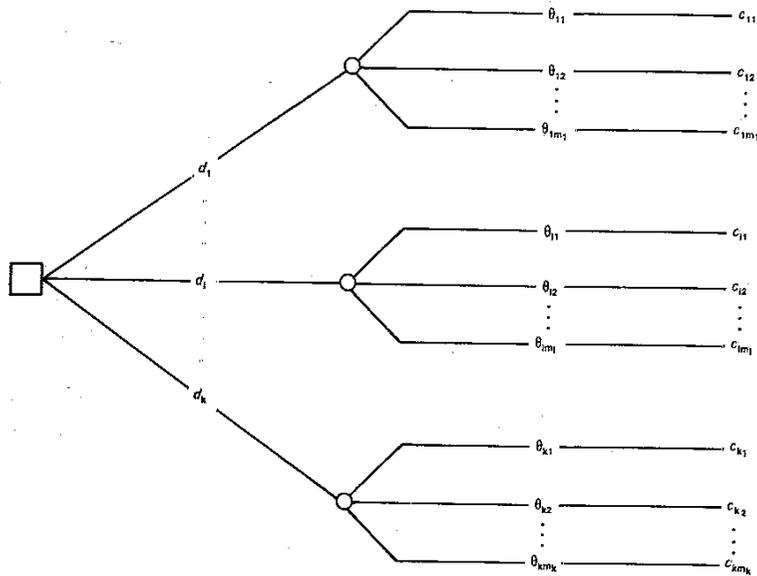
En principio, el espacio D puede contener infinitas alternativas. Sin embargo, en la mayor parte de las aplicaciones, D es un conjunto finito, lo que justifica que nos limitemos a considerar este caso cuando ello dé lugar a una simplificación matemática notable.

Determinar la mejor de un conjunto de alternativas sería, en principio, inmediato si tuviésemos información completa sobre las consecuencias de cada una de ellas. El vendedor de periódicos que debe decidir sobre el número de ejemplares con que se queda no tendría problema si supiese el número que podría vender. El médico que ante un caso determinado duda entre un tratamiento médico y uno quirúrgico no vacilaría si conociese las causas y el desarrollo de la afección. La principal dificultad con que uno se encuentra al plantearse un problema de decisión consiste en la falta de información sobre lo que sucederá según se actúe de una u otra manera. El problema general de decisión se plantea, pues, en *ambiente de incertidumbre*.

Existen situaciones en las que se tiene información completa y, sin embargo, es difícil tomar la decisión correcta, pero en estos casos la dificultad es de tipo técnico, no conceptual. Así, por ejemplo, a pesar de disponer de toda la información relevante, es difícil decidir cual es la mejor estrategia en un momento determinado de una partida de ajedrez o determinar la dieta más barata que cumple ciertos requisitos de nutrición. Sin embargo, la dificultad en estos casos es sólo de tipo técnico: el enorme número de estrategias posibles en el primer caso y el problema matemático de encontrar un máximo condicionado en el segundo, pero no aparecen dudas sobre el *criterio de decisión* que debe adoptarse. En este libro no consideraremos tales dificultades técnicas: supondremos que en presencia de información completa siempre puede elegirse la mejor de un conjunto de alternativa. Nos ocuparemos en su lugar del *proceso lógico de decisión* en ambiente de incertidumbre, es decir del método a seguir para tomar decisiones cuando *no* se dispone de toda la información que se juzga relevante.

Puesto que la dificultad esencial en un problema de decisión reside en las incertidumbres presentes en la situación es necesario considerar éstas con cuidado e introducirlas en la teoría. Así, una vez determinado el espacio de decisiones, habrá que considerar para cada una de las decisiones posibles

el conjunto de *sucesos inciertos* que determinan sus eventuales *consecuencias*. Esquemáticamente, la situación, en el caso de un número finito de alternativas y un número finito de sucesos inciertos puede representarse mediante un *árbol de decisión* de la forma



donde $D = \{d_1, d_2, \dots, d_k\}$ es el espacio de las decisiones y $\Theta_i = \{\theta_{i1}, \theta_{i2}, \dots, \theta_{im_i}\}$ es el conjunto de los m_i sucesos inciertos cuya eventual ocurrencia afecta al resultado de tomar la decisión d_i , de forma que si se toma la decisión d_i y sucede θ_{ij} se obtiene la consecuencia c_{ij}

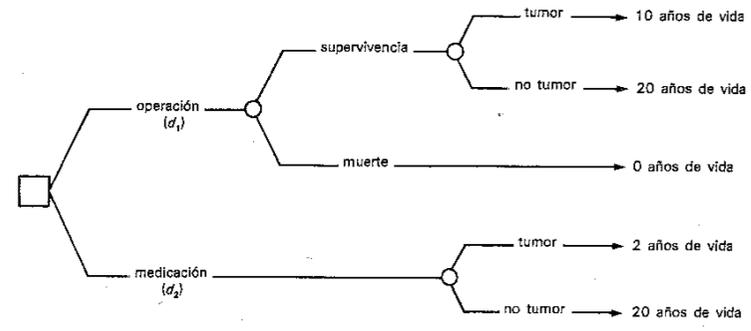
El diagrama empieza en un *nodo de decisión* representado por un cuadrado. Cualquiera que sea la decisión elegida se llega a un *nodo aleatorio*, representado por un círculo, sobre el que el decisor, no tiene control alguno. Las ramas de los nodos aleatorios pueden subdividirse, dando lugar a nuevos nodos aleatorios, si la relevancia de determinados sucesos inciertos dependen de que ocurran o no algunos de ellos. El árbol debe constuirse de forma que los sucesos inciertos a que da lugar cada uno de los nodos aleatorios sean mutuamente excluyentes y constituyan un conjunto exhaustivo. Así, por ejemplo, con referencia a la figura, si se toma la decisión d_1 tiene que ocurrir uno y sólo uno de los m_1 sucesos $\{\theta_{11}, \theta_{12}, \dots, \theta_{1m_1}\}$. Como en el caso de las

decisiones, siempre puede conseguirse que los sucesos inciertos correspondientes a un nodo aleatorio sean mutuamente excluyentes mediante la construcción de los apropiados espacios producto, pero la exhaustividad no es fácil de garantizar; la construcción de espacios de sucesos inciertos que contemplen todas las eventualidades relevantes suele exigir un conocimiento profundo del área de aplicación

Ejemplo 2.1.1. Oportunidad de una operación

Un médico debe decidir si realizar una peligrosa operación a una persona que se cree puede tener un tumor, o recurrir a una determinada medicación. Si el paciente no tiene el tumor su esperanza de vida se estima en 20 años. Si lo tiene, se opera y sobrevive a la operación, en 10 años, y si tiene el tumor y no se opera, sólo se le dan 2 años de vida. Construir el correspondiente árbol de decisión

El espacio de decisiones tiene claramente solo dos elementos $d_1 =$ operar y $d_2 =$ medicar. Si se realiza la operación el paciente puede sobrevivir o morir en ella y, si sobrevive, la consecuencia final depende de que tenga o no tenga el tumor. Si no se opera no puede morir en la operación, pero la consecuencia final dependerá de nuevo de que el paciente tenga o no el tumor. El correspondiente árbol de decisión es, pues,



La mayoría de los problemas de decisión tienen, aparentemente, una estructura más compleja que la del problema de decisión que hemos descrito. Por ejemplo, se puede plantear inicialmente si realizar o no un experimento y, en caso afirmativo, tratar de decidir cual es la acción más adecuada según el resultado del experimento. Podemos, asimismo, considerar problemas secuenciales de decisión constituidos por la yuxtaposición de problemas como el anterior. Sin embargo, como veremos más adelante, tales problemas complejos de decisión pueden ser resueltos analizando sucesivamente cada uno de los subproblemas, como el descrito, que lo constituyen, por lo que bastará que nos ocupemos de resolver éste último.

Frecuentemente, el conjunto de sucesos inciertos relevantes es el mismo cualquiera que sea la decisión que se tome, es decir $\Theta_i = \{\theta_{i1}, \theta_{i2}, \dots, \theta_{im_i}\} = \{\theta_1, \theta_2, \dots, \theta_m\} = \Theta$ para todo i . Si, además, sólo hay un número finito de alternativas y de sucesos inciertos, entonces el problema de decisión puede representarse también mediante una *tabla de decisión* de la forma

	θ_1	\dots	θ_j	\dots	θ_m
d_1	c_{11}	\dots	c_{1j}	\dots	c_{1m}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
d_i	c_{i1}	\dots	c_{ij}	\dots	c_{im}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
d_k	c_{k1}	\dots	c_{kj}	\dots	c_{km}

donde $D = \{d_1, d_2, \dots, d_k\}$ es el conjunto de decisiones posibles, $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$ el conjunto de sucesos inciertos relevantes, cualquiera que sea la decisión tomada, y c_{ij} la consecuencia de tomar la decisión d_i y que suceda θ_j .

Ejemplo 2.1.2. Elección de un medio de transporte

Se dispone de un automóvil y de una motocicleta para ir a resolver un asunto en el centro de la ciudad. La moto es más rápida, sobre todo si se producen atascos, y puede aparcarse con facilidad en el mismo centro, pero resulta incómoda si hace mal tiempo. El coche, aunque más cómodo, consume más gasolina y hay que dejarlo en un aparcamiento a cierta distancia del lugar de destino para caminar a continuación. Suponemos que no se dispone de una línea adecuada de transporte público y que se considera excesivo el coste de un taxi. Construir la correspondiente tabla de decisión.

En estas condiciones, el espacio de decisiones alternativas es $D = \{d_1, d_2, d_3\}$ con

- d_1 = ir en motocicleta
- d_2 = ir en automóvil
- d_3 = ir a pie

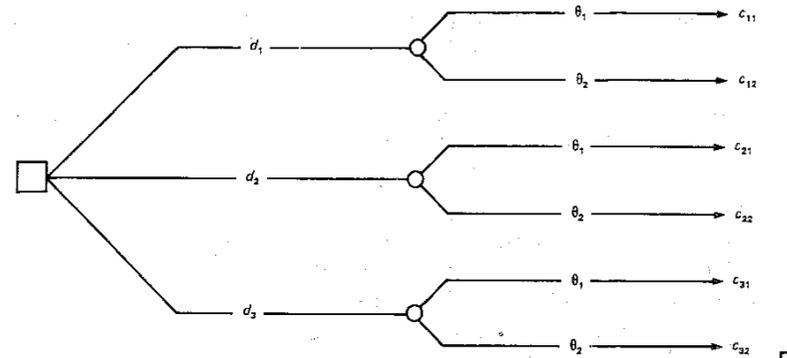
Un suceso incierto del que claramente dependen las consecuencias de la decisión es el suceso

θ_1 = empieza a llover antes del regreso

Si no se considera relevante ningún otro suceso incierto, el espacio de sucesos inciertos mutuamente excluyentes es simplemente $\Theta = \{\theta_1, \theta_2\}$, donde θ_2 es la negación de θ_1 , y el conjunto de consecuencias puede describirse así

	$\theta_1 = \text{lueve}$	$\theta_2 = \text{no llueve}$
$d_1 = \text{moto}$	c_{11} { No se pierde tiempo Trayecto en moto muy incómodo Pequeño coste	c_{12} { No se pierde tiempo Agradable paseo en moto Pequeño coste
$d_2 = \text{coche}$	c_{21} { Se pierde bastante tiempo Conducción por ciudad con mal tiempo Coste moderado	c_{22} { Se pierde bastante tiempo Conducción por ciudad con buen tiempo Coste moderado
$d_3 = \text{a pie}$	c_{31} { Se pierde mucho tiempo Caminata en mal tiempo No hay coste	c_{32} { Se pierde mucho tiempo Caminata con buen tiempo No hay coste

Naturalmente, el mismo problema de decisión puede representarse mediante un árbol de decisión en la forma



donde los c_{ij} tienen el significado dado en la tabla

Cualquier árbol de decisión, incluso cuando los conjuntos de sucesos inciertos relevantes son distintos para cada decisión d_i , puede representarse en forma de tabla mediante el artificio de considerar como conjunto Θ de sucesos inciertos el espacio producto de los conjuntos correspondientes a cada decisión d_i , pero este procedimiento oscurece y dificulta la solución del problema

En resumen, nuestro modelo de problema de decisión exige la especificación del espacio D de las decisiones alternativas y de los conjuntos Θ_i de sucesos inciertos mutuamente excluyentes que condicionan las consecuencias de cada una de las decisiones. El problema de decisión consiste entonces en elegir una decisión d_i del conjunto D sin saber cual de los sucesos θ_j de Θ_i tendrá lugar.

En la próxima sección describiremos una solución intuitiva a un problema de decisión; más adelante demostraremos que esa solución es la única compatible con ciertos principios elementales de comportamiento coherente.

2.2. Solución intuitiva a un problema de decisión

Aunque los sucesos que componen cada Θ_i son inciertos, en el sentido de que no sabemos cual de ellos tendrá lugar, no nos resultan, en general, igualmente verosímiles. En efecto, aunque no se disponga de la información suficiente para determinar cual de ellos tendrá lugar, típicamente se dispone de cierta información que hace unos elementos de cada Θ_i más verosímiles que otros. Así, aunque no es posible anticipar con seguridad los años de vida que le quedan a una persona sana que acaba de cumplir los 40, nos parece más verosímil que viva otros 10 años a que muera el mes próximo o a que llegue a centenario.

Nuestro primer objetivo sería precisar de forma cuantitativa el contenido de este tipo de información incompleta. Una forma de hacerlo sería asignar a cada suceso incierto un número que midiese la verosimilitud que se le atribuye, de forma que a los sucesos más verosímiles se les haga corresponder un número mayor.

Desde hace varios siglos, la incertidumbre existente sobre la ocurrencia de *algunas* clases de sucesos ha venido midiéndose por un número entre 0 y 1 al que se ha llamado su *probabilidad*.

Para hacerlo, la *unidad* de probabilidad correspondiente a un suceso *cierto* se *distribuye* entre una clase de sucesos mutuamente excluyentes de forma que el número asignado a un determinado suceso —su probabilidad— es tanto mayor cuanto más verosímil se juzga la posibilidad de que tenga lugar.

Supóngase, por ejemplo, que se lanza una moneda al aire sobre una superficie lisa. Sólo pueden tener lugar dos sucesos inciertos mutuamente excluyentes: cara o cruz. La mayoría de nosotros diríamos que ambos sucesos nos parecen igualmente verosímiles de forma que si debemos *distribuir* entre ellos la unidad de probabilidad les asignaríamos $1/2$ a cada uno. De forma similar, si lanzamos un dado sólo pueden tener lugar seis sucesos mutuamente excluyentes que de nuevo consideraríamos igualmente verosímiles, por lo que si debemos distribuir entre ellos la unidad de probabilidad les asignaríamos $1/6$ a cada uno. En general, si *considerásemos igualmente verosímiles* a n sucesos mutuamente excluyentes les asignaríamos a cada uno de ellos una probabilidad $1/n$.

Existe otra clase de sucesos a los que resulta fácil asignar una probabilidad aunque no presenten las propiedades de simetría características de los juegos de azar. Se trata de aquellos sucesos que es posible observar repetidamente en *idénticas* condiciones. En efecto, consideremos una situación B que puede o no dar lugar a un suceso A , y supongamos que tal situación puede repetirse indefinidamente. Es un hecho empírico que, frecuentemente, el cociente m/n entre el número de veces m que tiene lugar el suceso A y

el número total n de observaciones se estabiliza y parece tender a un límite. Parece razonable pensar que, en *ausencia de otra información relevante*, tal límite mediría en una escala de 0 a 1 la verosimilitud que se concede al suceso A en las condiciones descritas por B . Se sabe, por ejemplo, que aproximadamente el 51 % de los nacimientos humanos dan lugar a un varón. Es por tanto un hecho empírico que los nacimientos humanos tienden a *distribuirse* en un 51 % de varones y un 49 % de hembras. En consecuencia, en ausencia de otra información relevante, parece razonable distribuir en 0,51 y 0,49 la unidad de probabilidad y describir con tales números las probabilidades de que una mujer embarazada dé a luz, respectivamente, a un varón o a una hembra.

La existencia de simetrías o de datos empíricos sobre frecuencias relativas es ciertamente *útil* para precisar cuantitativamente nuestra incertidumbre sobre un determinado suceso pero no es en modo alguno *necesaria*. En efecto, no se comienza una escalada si parece *probable* que vaya a haber tormenta ni se empieza una formación profesional si no parece *probable* encontrar trabajo al terminarla; continuamente asignamos probabilidades, aunque sea de forma completamente inconsciente, a todo tipo de sucesos. Debe advertirse, además, que la mayoría de sucesos relevantes a un problema de decisión no presentan simetrías ni son repetibles, de forma que un concepto de probabilidad restringido a tales clases de sucesos sería inevitablemente insuficiente. Así, por ejemplo, las consecuencias de muchas decisiones pueden depender de un cambio político en el país. Para tomar razonablemente tales decisiones resulta, pues, necesario valorar la verosimilitud de tal cambio. Es obvio que las crisis políticas no presentan simetrías ni son repetibles: resulta clara la necesidad de un concepto de probabilidad que pueda aplicarse a cualquier tipo de suceso.

Para nosotros, la *probabilidad* de un suceso A en una situación H , que representaremos por $p(A|H)$, será una medida sobre una escala $[0, 1]$ de la verosimilitud que se concede al suceso A en las condiciones descritas por H , esto es, una medida del *grado de creencia* en A que nos sugiere la información contenida en H . En un extremo, si dado H se está seguro de que A tendrá lugar, $p(A|H) = 1$; en el otro extremo, si dado H se está convencido de que A no sucederá, $p(A|H) = 0$. Otros valores en el intervalo $(0, 1)$ expresan grados de creencia intermedios. La probabilidad asignada a un suceso es siempre *condicional* a la información que se posee sobre él: no existen probabilidades «absolutas». Sin embargo, con objeto de simplificar la notación, la condición H será omitida cuando ello no dé lugar a confusión, escribiéndose $p(A)$ en lugar de $p(A|H)$.

Volviendo al problema de decisión, la información que el decisor posee sobre la verosimilitud de los distintos sucesos inciertos de cada Θ_i en el momento de tomar la decisión podría pues ser cuantificada *distribuyendo* la unidad de probabilidad, para cada d_i , entre los sucesos $\{\theta_{i1}, \theta_{i2}, \dots, \theta_{im}\}$ que

comprende Θ_i . Puesto que, por hipótesis, estos sucesos son mutuamente excluyentes, podríamos describir la información disponible sobre su verosimilitud, en las condiciones H en que debe tomarse la decisión, mediante un conjunto de números $\{p(\theta_{ij}|d_i, H), j = 1, 2, \dots, m_i\}$ para cada i , tales que

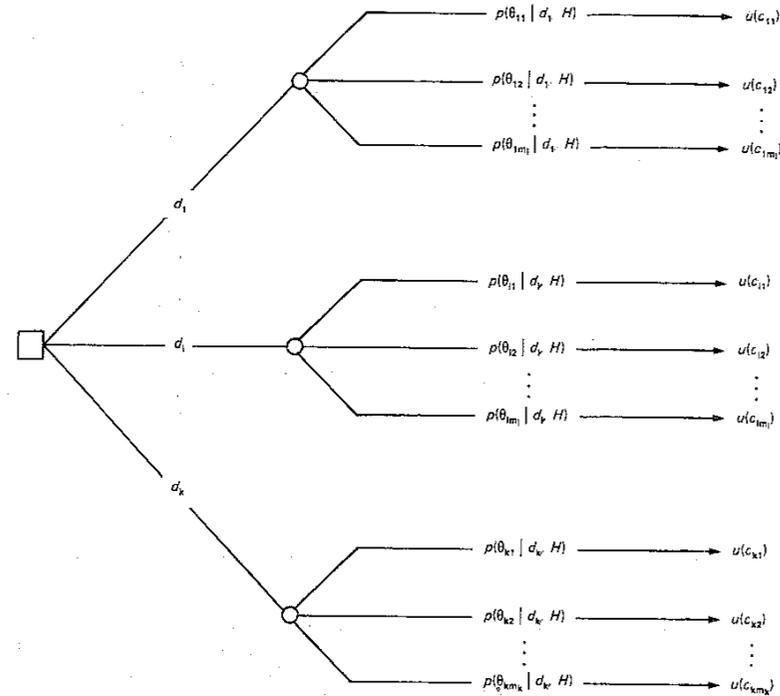
$$0 \leq p(\theta_{ij}|d_i, H) \leq 1, \quad \sum_{j=1}^{m_i} p(\theta_{ij}|d_i, H) = 1 \quad (1)$$

Si no hay confusión posible, omitiremos la condición H y escribiremos simplemente $p(\theta_{ij}|d_i)$ en lugar de $p(\theta_{ij}|d_i, H)$.

Consideremos un problema de decisión en el que tanto el espacio de decisiones como el de sucesos inciertos son finitos. Sea $D = \{d_1, d_2, \dots, d_k\}$ el conjunto de decisiones alternativas y $\Theta_i = \{\theta_{i1}, \theta_{i2}, \dots, \theta_{im_i}\}$ el de sucesos inciertos mutuamente excluyentes correspondientes a la decisión d_i . De acuerdo con la notación empleada en la sección anterior, llamaremos c_{ij} a la consecuencia de que suceda θ_{ij} cuando se ha adoptado la decisión d_i .

Obviamente, el decisor tendrá sus preferencias entre las distintas consecuencias. En principio, tales preferencias podrían ser cuantificadas asignando a cada una de las consecuencias c_{ij} un número $u(c_{ij})$ que midiese la *utilidad* que cada una de ellas tuviese para el decisor. Por ejemplo, podría asignarse el valor u_1 a la consecuencia más preferida, el valor u_0 a la menos preferida y valores intermedios $u(c_{ij})$ al resto de forma que si una consecuencia es preferida a otra le sea asignado un número mayor. La *función de utilidad* así construida mide las preferencias del decisor entre las posibles consecuencias de su decisión, de forma parecida a como la probabilidad antes descrita mide la verosimilitud que le merecen los posibles sucesos inciertos. Los conceptos duales de probabilidad y utilidad, que aquí se utilizan en su sentido coloquial, serán formalizados, respectivamente, en las Secciones 2.4 y 2.5.

Una vez especificadas las probabilidades $p(\theta_{ij}|d_i, H)$ que describen la verosimilitud de los sucesos inciertos y las utilidades $u(c_{ij})$ que describen las preferencias del decisor entre las posibles consecuencias, el problema planteado tiene ya una solución inmediata. En efecto, introduciendo en el árbol de decisión las probabilidades y las utilidades mencionadas, tendríamos



Así, si se toma la decisión d_i en las condiciones H se puede obtener utilidad $u(c_{i1})$ con probabilidad $p(\theta_{i1}|d_i, H)$, $u(c_{i2})$ con probabilidad $p(\theta_{i2}|d_i, H)$, ..., $u(c_{im_i})$ con probabilidad $p(\theta_{im_i}|d_i, H)$. Por lo tanto, la utilidad *media* de la decisión d_i , a la que llamaremos la *utilidad esperada* de d_i y representaremos por $u^*(d_i)$, vendrá dada por

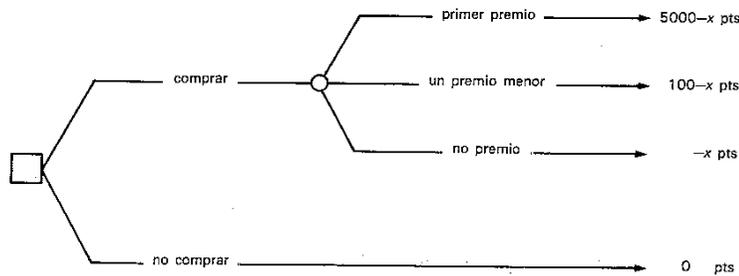
$$u^*(d_i) = \sum_{j=1}^{m_i} u(c_{ij}) p(\theta_{ij}|d_i, H) \quad (2)$$

Resulta natural elegir como decisión más razonable aquella que *maximiza la utilidad esperada*, esto es aquella que hace máxima la expresión (2) entre las k alternativas $\{d_1, d_2, \dots, d_k\}$. Este es el *criterio Bayes* para la toma de decisiones.

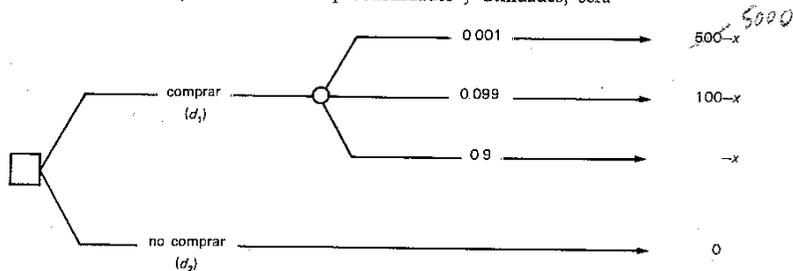
Ejemplo 2.2.1. Participación en una lotería

En una rifa que tiene mil números, se sortea un premio de 5.000 pts y se dan 100 a todos los billetes cuya última cifra coincide con la del primer premio. Suponiendo la utilidad del dinero proporcional a su cantidad, de terminar el precio máximo que debe pagarse por participar

Si representamos por x el coste de participar en la rifa, el árbol de decisión en términos de sucesos inciertos y consecuencias de



La probabilidad de obtener el primer premio es obviamente $1/1\,000 = 0,001$ y la de obtener un premio menor $(100 - 1)/1\,000 = 0,099$. En consecuencia, la probabilidad de quedarse sin premio es 0,9. Como suponemos las utilidades proporcionales al dinero, el árbol de decisión, en términos de probabilidades y utilidades, será



Usando (2), la utilidad esperada de comprar (d_1) será

$$u^*(d_1) = 0,001 (5\,000 - x) + 0,099 (100 - x) - 0,9 x = 14,9 - x$$

y esta será la decisión razonable si, y sólo si, $u^*(d_1) > u^*(d_2) = 0$ esto es si $14,9 - x > 0$ y por lo tanto si $x < 14,9$; el precio máximo que debe pagarse por un número de la rifa es 14,9 pts

Si el conjunto de sucesos inciertos relevantes es el mismo cualquiera que

sea la decisión que se tome, de forma que $\Theta_i = \Theta = \{\theta_1, \theta_2, \dots, \theta_m\}$ para todo i , es más cómoda la representación en forma de tabla. Así, si introducimos en la tabla de decisión las probabilidades $\{p(\theta_j|H), j = 1, 2, \dots, m\}$ que describen la verosimilitud de los sucesos inciertos en las condiciones H en que hay que decidir, de forma que

$$0 \leq p(\theta_j|H) \leq 1 \quad \text{y} \quad \sum_{j=1}^m p(\theta_j|H) = 1 \quad (3)$$

y las utilidades $u(c_{ij})$ que describen las preferencias del decisor, obtendremos

	$p(\theta_1 H)$	$p(\theta_2 H)$	\dots	$p(\theta_m H)$
d_1	$u(c_{11})$	$u(c_{12})$	\dots	$u(c_{1m})$
\dots	\dots	\dots	\dots	\dots
d_i	$u(c_{i1})$	$u(c_{i2})$	\dots	$u(c_{im})$
\dots	\dots	\dots	\dots	\dots
d_k	$u(c_{k1})$	$u(c_{k2})$	\dots	$u(c_{km})$

Si se toma la decisión d_i en las condiciones H se puede obtener una utilidad $u(c_{i1})$ con probabilidad $p(\theta_1|H)$, $u(c_{i2})$ con probabilidad $p(\theta_2|H)$, ..., $u(c_{ik})$ con probabilidad $p(\theta_k|H)$. En consecuencia, la utilidad esperada de la decisión d_i es

$$u^*(d_i) = \sum_{j=1}^m u(c_{ij}) p(\theta_j|H) \quad (4)$$

Obviamente, las ecuaciones (3) y (4) son un caso particular de (1) y (2). La decisión más razonable, la que maximiza la utilidad esperada, o decisión Bayes, es aquella que maximiza la expresión (4).

Ejemplo 2.2.2. Elección de un medio de transporte (cont)

Considérese de nuevo el problema del Ejemplo 2.1.2; especifíquese una función de utilidad y determínese, en función de la probabilidad p de que llueva antes del regreso, cual es la decisión más razonable.

Dado un conjunto de consecuencias, parece razonable suponer que puede encontrarse una consecuencia c^* , perteneciente o no a ese conjunto, que sea preferible o igualmente deseable a cualquiera de las que lo integran, y otra consecuencia c_* , que no es estrictamente preferible a ninguna de ellas. De esta forma, para cualquier consecuencia c del mencionado conjunto, tendremos $c_* \leq c \leq c^*$. Para evitar el caso trivial de que todas las consecuencias sean igualmente deseables, supondremos además que $c_* < c^*$.

POSTULADO C1. Para todo par de opciones l_1 y l_2 , es cierta una de las tres relaciones $l_1 < l_2$, $l_1 > l_2$ o $l_1 \sim l_2$. Además, es posible encontrar dos consecuencias c^* y c_* tales que $c^* > c_*$ y que para toda consecuencia c , $c_* \leq c \leq c^*$.

La hipótesis de comparabilidad es esencial para el resultado que nos proponemos demostrar. Conviene, pues, considerarla con cierto detalle. Puesto que, según hemos visto, las decisiones son casos particulares de opciones, el postulado C1 equivale a suponer que el problema de decisión en ambiente de incertidumbre tiene solución, esto es que no existen dos decisiones incomparables; que el decisor siempre prefiere una de ellas, o las considera igualmente deseables. El postulado de comparabilidad refleja el hecho observable de que en la práctica se comparan opciones de muy distinto tipo, forzados por la necesidad de actuar. Así, por ejemplo, se argumenta a menudo que el valor del tiempo libre no puede medirse; sin embargo, cada uno de nosotros le asigna un valor concreto en el momento que decide si aceptar o no un trabajo extraordinario.

La comparabilidad de opciones implica en particular la comparabilidad de consecuencias y de verosimilitudes. Ya hemos visto que las consecuencias son casos particulares de opciones. Además, comparar la opción $l_1 = \{c^*|A, c_*|\bar{A}\}$ con $l_2 = \{c^*|B, c_*|\bar{B}\}$, donde \bar{A} , \bar{B} son, respectivamente, los sucesos complementarios de A y B , equivale a comparar las verosimilitudes de A y B . Obviamente l_1 será preferible a l_2 si, y solo si, A es más verosímil que B . Sin embargo, hay quien ha sugerido que existen dos tipos de sucesos: unos cuya verosimilitud puede ser fácilmente descrita en términos cuantitativos y otros que no admiten tal descripción. Entre los primeros estarían los sucesos relacionados con los juegos de azar y aquellos de los que se conoce empíricamente su frecuencia relativa. Entre los segundos se contarían sucesos no repetibles tales como los eventuales resultados de una elección política. Los primeros serían sucesos *estadísticos*, cuyas verosimilitudes podrían ser comparadas entre sí, los segundos serían sucesos *no estadísticos* cuyas verosimilitudes no podrían ser valoradas y comparadas con las de otros sucesos. Sin embargo, tal clasificación es ilusoria. En primer lugar, es fácil encontrar sucesos que se resisten a ser clasificados de esta manera. Considérese, por ejemplo, el

suceso «Jorge aprobará en el primer intento el examen de conducir». Ciertamente, las estadísticas sobre la proporción de personas que aprueban al primer intento pueden proporcionar información relevante sobre la verosimilitud del suceso que nos ocupa, que parecería, por tanto, ser un suceso estadístico. Sin embargo, también es relevante nuestra información sobre el carácter, los reflejos y la habilidad mecánica de Jorge, y esto nos induciría a pensar que el suceso planteado *no* es estadístico. En segundo lugar, sucesos como «muerte en accidente de carretera» que son ciertamente estadísticos para una casa de seguros no lo son necesariamente para un individuo. ¿Hasta qué punto es relevante la proporción de conductores que perecen en la carretera para determinar la probabilidad de que *usted*, lector, tenga un accidente mortal? Si los sucesos no pueden ser clasificados en comparables e incomparables, y ciertamente efectuamos algunas comparaciones, parece razonable sugerir que con un esfuerzo adecuado siempre podríamos decidir cual de dos sucesos específicos nos parece más verosímil o si ambos nos parecen igualmente verosímiles.

(ii) TRANSITIVIDAD

Cualquiera de nosotros se encontraría probablemente incómodo si le puntualizasen que en un momento de su razonamiento había afirmado que la opción l_1 era preferible a la l_2 y la l_2 a la l_3 , pero que acababa de afirmar que l_3 era preferible a l_1 . Postulamos pues que las preferencias deben ser transitivas.

POSTULADO C2. Si $l_1 > l_2$ y $l_2 > l_3$, entonces $l_1 > l_3$. Análogamente si $l_1 \sim l_2$ y $l_2 \sim l_3$, entonces $l_1 \sim l_3$.

Resulta fácil, por reducción al absurdo, defender el postulado de transitividad. En efecto, si el decisor prefiere l_1 a l_2 y l_2 a l_3 , debería estar dispuesto a pagar una cierta cantidad x por pasar de la situación l_3 a la situación l_2 , y otra cantidad y por pasar de la l_2 a la l_1 . Pero si además prefiere l_3 a l_1 , también debería estar dispuesto a pagar una cierta cantidad z por sustituir la situación l_1 por la l_3 . En consecuencia, nuestro intransitivo decisor habría vuelto a la situación original, l_3 , tras haber pagado $x + y + z$, y el proceso podría repetirse; un decisor intransitivo sería pues una especie de máquina perpetua de regalar dinero. Un argumento totalmente análogo puede utilizarse para justificar que si la opción l_1 es equivalente a la l_2 y la l_2 a la l_3 , entonces la opción l_1 debe ser equivalente a la l_3 .

(iii) SUSTITUCIÓN Y DOMINANCIA

Claramente, la equivalencia entre opciones puede ser establecida por partes. Así por ejemplo, si la opción l_1 se juzga equivalente a la opción l_2 en

los días laborables, y también en los festivos, entonces las opciones l_1 y l_2 deben siempre ser juzgadas equivalentes. Formalmente si representamos por \bar{A} el suceso complementario de A ,

POSTULADO C3 Si $l_1 > l_2$ cuando sucede A y $l_1 > l_2$ cuando sucede \bar{A} , entonces $l_1 > l_2$. Análogamente, si $l_1 \sim l_2$ cuando sucede A y $l_1 \sim l_2$ cuando sucede \bar{A} , entonces $l_1 \sim l_2$.

En general, si dos opciones $l_1 = \{c_{11}|A_1, \dots, c_{1k}|A_k\}$ y $l_2 = \{c_{21}|A_1, \dots, c_{2k}|A_k\}$ tienen los mismos sucesos inciertos A_i y las consecuencias de la primera dominan a las de la segunda, de forma que $c_{1i} \geq c_{2i}$ para todo i , entonces $l_1 \geq l_2$; si, además, $c_{1i} > c_{2i}$ para algún valor de i entonces $l_1 > l_2$.

En virtud del principio de sustitución, en una opción puede reemplazarse una consecuencia por otra opción equivalente a ella. En efecto, si $c_1 \sim l_1$ la opción $l_2 = \{c_1|A, c_2|\bar{A}\}$ es equivalente a $l_3 = \{l_1|A, c_2|\bar{A}\}$. Basta observar que si sucede A , l_2 da lugar a c_1 y l_3 a l_1 y puesto que por hipótesis $c_1 \sim l_1$ tenemos que, si sucede A , $l_2 \sim l_3$. Si sucede \bar{A} , ambas opciones dan lugar a c_2 y por lo tanto $l_2 \sim l_3$ si sucede \bar{A} . Consecuentemente, en virtud del Postulado C3, $l_2 \sim l_3$.

(iv) SUCESOS DE REFERENCIA

Hemos anticipado que para poder tomar decisiones de forma razonable es necesario medir la información y las preferencias del decisor expresándolas de forma cuantitativa. Para poder medir es necesaria una *unidad de medida*. Con este objeto, postularemos que es posible imaginar métodos para elegir puntos al azar en el cuadrado unidad, de forma que resulte igualmente verosímil elegir cualquier punto, y los utilizaremos para construir con ellos una familia de sucesos que nos sirvan como referencia, como unidad de medida.

POSTULADO C4. El decisor puede concebir un procedimiento de generar un punto aleatorio z en el cuadrado unidad, esto es, un número $z = (x, y)$, $0 \leq x \leq 1$, $0 \leq y \leq 1$ tal que para cualquier par de regiones R_1, R_2 del cuadrado unidad el suceso $\{z \in R_1\}$ le resulta menos verosímil que el suceso $\{z \in R_2\}$ si, y solo si, el área de R_1 es menor que la de R_2 .

Es fácil imaginar métodos de generar puntos aleatorios en el cuadrado unidad. Por ejemplo, haciendo rodar una ruleta de circunferencia unidad y anotando la longitud del arco, en sentido positivo, que separa un origen prefijado del punto señalado por la aguja, se obtendrá un número aleatorio x en $[0, 1]$. Repitiendo el proceso se obtendrá otro número aleatorio y en $[0, 1]$. El punto $z = (x, y)$ determinado por las coordenadas x, y sería entonces un

punto aleatorio en el cuadrado unidad. Obsérvese que el Postulado C4 se limita a afirmar que el decisor puede *imaginar* un procedimiento de generar puntos en el cuadrado unidad que él considere aleatorios. No se exige la construcción física de tal procedimiento ni su carácter «objetivo».

Para simplificar la notación y siempre que no se preste a confusión, denotaremos por R tanto una *región* del cuadrado unidad como el *suceso* de que un punto aleatorio, en el sentido del Postulado C4 se sitúe en dicha región. Donataremos por \bar{R} al suceso complementario de R en el cuadrado unidad.

El Postulado C4 nos permite construir una batería de opciones con las que medir, por comparación, la deseabilidad de todas las demás. Específicamente, consideraremos opciones de la forma

$$\{c^*|R, c_*|\bar{R}\}, \quad R \subset [0, 1]^2$$

esto es situaciones en las que se obtiene la mejor consecuencia posible c^* si sucede $\{z \in R\}$ y la peor posible c_* si tiene lugar el suceso complementario $\{z \notin R\}$, donde z es un punto aleatorio del cuadrado unidad.

Ejemplo 2.3.1. Opciones económicas alternativas

Considérense los sucesos

A_1 : el índice de inflación de este año será menor del 15 %.

A_2 : habrá elecciones generales dentro del próximo año, y considérense las opciones alternativas a que se reducen determinadas decisiones, que vienen dadas por

$$l_1 = \{1000|A_1, \quad -800|\bar{A}_1\}$$

$$l_2 = \{500|A_2, \quad -100|\bar{A}_2\}$$

$$l_3 = \{200|A_1 \cap A_2, \quad 100|A_1 \cap \bar{A}_2, \quad -150|\bar{A}_1\}$$

$$l_4 = \{400|A_1 \cap A_2, \quad 300|\bar{A}_1 \cap A_2, \quad -350|\bar{A}_2\}$$

donde las consecuencias se han expresado en miles de pesetas. Especificar preferencias entre los seis posibles pares de opciones que satisfagan los principios de coherencia

Obviamente, no existe una solución única a este problema. Una solución que satisfaga los principios de coherencia, cuando A_1 , no parece muy verosímil pero A_2 sí lo parece, es

$$\begin{aligned} l_1 < l_2 & \quad l_1 < l_3 & \quad l_1 < l_4 \\ l_2 > l_3 & \quad l_2 > l_4 \\ l_3 < l_4 \end{aligned}$$

que conjuntamente implican $l_1 < l_3 < l_4 < l_2$

2.4. Probabilidad como grado de creencia

En la Sección 2.2 comentamos la necesidad de determinar la probabilidad asignada por el decisor a los distintos sucesos inciertos que puedan influir en las las consecuencias de sus decisiones.

Históricamente, el concepto de probabilidad surgió con el estudio de los juegos de azar, en los que se dan ciertas simetrías que permiten concebir la probabilidad de un suceso como el cociente entre el número de casos en que puede darse y el número de casos totales, cuando todos los casos se juzgan igualmente verosímiles y mutuamente excluyentes. Así, puesto que al lanzar un dado suelen juzgarse las seis caras igualmente verosímiles, la probabilidad de obtener un número par sería $3/6, = 1/2$.

Más tarde, con el estudio de los problemas planteados por las compañías de seguros, apareció el concepto de probabilidad de un suceso como el límite a que tendería la frecuencia relativa con que ese suceso se presentaría si se repitiese indefinidamente la misma situación en idénticas condiciones. Así, puesto que aproximadamente el 49 % de los nacimientos humanos dan lugar a niñas, la probabilidad de que una mujer embarazada tenga una niña sería 0,49

Obviamente, la mayor parte de los sucesos inciertos que intervienen en un problema de decisión no presentan simetrías ni pueden ser repetidos indefinidamente en idénticas condiciones. Así, por ejemplo, para decidir si debe ser utilizado un nuevo fármaco en un determinado paciente, hay que estimar la probabilidad de que este nuevo fármaco le sea eficaz y no le produzca trastornos secundarios; sin embargo, no existen simetrías ni experiencia previa en idénticas condiciones que permitan especificar la probabilidad buscada.

Aunque pueden servir para cuantitativizar la verosimilitud de determinados sucesos, los conceptos de simetría y frecuencia relativa no sirven, como algunos han pretendido, para *definir* el concepto de probabilidad. Además de referirse a clases de sucesos demasiado restringidas para ser útiles en un problema de decisión, tales definiciones resultarían necesariamente circulares. En efecto, una definición de *probabilidad* no puede basarse en la existencia de sucesos igualmente *probables* como la definición clásica «casos favorables/casos posibles» exige. Por otra parte, al afirmar la existencia de un «límite»

de las frecuencias relativas del tipo $p = \lim m/n$ se hace referencia a un hecho empírico, no a un límite matemático; en realidad, incluso para valores enormes de n , no es *imposible* que m/n esté lejos de p , es meramente *improbable*. El intento de definición resulta así circular. Sin embargo, los postulados de coherencia que hemos descrito permiten definir la probabilidad de un suceso cualquiera. Esto se consigue comparando el suceso en cuestión con el suceso de que un punto aleatorio se sitúe en una determinada región del cuadrado unidad y eligiendo la región de forma que, para el decisor, ambos sucesos sean igualmente verosímiles.

DEFINICIÓN 2.4.1. La probabilidad de un suceso E en las condiciones H , que denotaremos por $p(E|H)$, es igual al área de una región R del cuadrado unidad elegida de forma que las opciones $l_1 = \{c^*|E, c_*|\bar{E}\}$ y $l_2 = \{c^*|R, c_*|\bar{R}\}$ sean igualmente deseables en las condiciones H .

En la definición \bar{E}, \bar{R} son, naturalmente, los sucesos complementarios de E, R y las consecuencias c^*, c_* las que aparecen definidas en el Postulado C1. Es fácil demostrar que la probabilidad de cualquier suceso queda así bien definida, esto es, que siempre existe una región del cuadrado unidad que cumple la condición exigida; que si dos regiones distintas la cumplen deben tener la misma área; y que la probabilidad asignada es independiente de las consecuencias de referencia c^*, c_* que se hayan elegido.

En efecto, comparando la desabilidad de $l = \{c^*|E, c_*|\bar{E}\}$ en las condiciones H con los elementos de una sucesión de opciones de la forma $l(x) = \{c^*|R(x), c_*|\bar{R}(x)\}$, donde $R(x)$ es el suceso de que un punto aleatorio se sitúe en una determinada región del cuadrado unidad de área x , tendremos necesariamente que $l(0) \leq l_1 \leq l(1)$ y, dada la continuidad de los números reales, deberá existir un x tal que $l = l(x)$, con lo que $p(E|H) = x$. Además, en virtud de la transitividad y de la definición de experimento auxiliar, si dos regiones distintas satisfacen las condiciones de la Definición 2.4.1, deben tener la misma área. Finalmente, el principio de sustituibilidad permite comprobar que el valor de $p(E|H)$ no se altera si se modifican las consecuencias de referencia.

La medida de probabilidad que hemos definido tiene las propiedades clásicamente atribuidas a las probabilidades. En efecto,

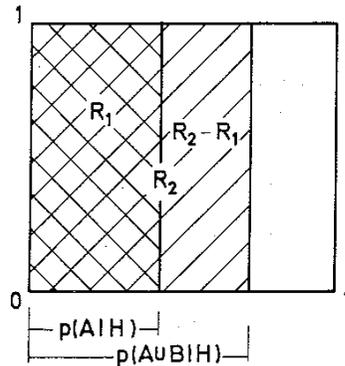
TEOREMA 2.4.1. Cualquiera que sean las condiciones de referencia H la medida de probabilidad verifica

- (i) Para todo suceso A , $0 \leq p(A|H) \leq 1$ y $p(H|H) = 1$
- (ii) Si A y B son dos sucesos incompatibles dado H ,
 $p(A \cup B|H) = p(A|H) + p(B|H)$.
- (iii) Para todo par de sucesos A y B ,
 $p(A \cap B|H) = p(A|H) \cdot p(B|A, H) = p(B|H) \cdot p(A|B, H)$.

Demostración

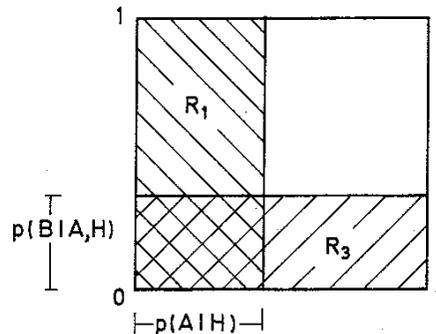
(i) Por construcción, tratándose de áreas de un subconjunto del cuadrado unidad todas las probabilidades están comprendidas entre 0 y 1. Además, dado H , el suceso H es un suceso cierto, igualmente verosímil por tanto a que un punto aleatorio se sitúe en el cuadrado unidad. Su probabilidad es, pues, el área del cuadrado unidad, es decir uno.

(ii) Sean R_1 y R_2 dos regiones del cuadrado unidad, tales que en las condiciones H , $R_1 \sim A$ y $R_2 \sim A \cup B$ (véase figura)



Entonces, en las condiciones H , $B \sim R_2 - R_1$. En efecto, si fuese $B < R_2 - R_1$, entonces, puesto que tanto A y B como R_1 y $R_2 - R_1$ son incompatibles, y además $A \sim R_1$, tendríamos que $A \cup B < R_1 \cup (R_2 - R_1) = R_2$, lo que contradice la definición de R_2 . De forma análoga, tampoco es posible que $B > R_2 - R_1$. Por tanto, en virtud de la comparabilidad, $B \sim R_2 - R_1$, en las condiciones H , y por tanto $p(B|H)$ será el área de $R_2 - R_1$, esto es $p(A \cup B|H) - p(A|H)$, como queríamos demostrar.

(iii) Sean R_1 y R_3 dos regiones del cuadrado unidad tales que, en las condiciones H , $R_1 \sim A$ y $R_3 \sim B|A$. Tales regiones pueden ser escogidas de la forma descrita en la figura



Claramente, dado H , $A \cap B \sim R_1 \cap R_3$, puesto que el suceso $A \cap B$ es equivalente

al suceso A seguido del $B|A$ y esto, en las condiciones H , es igualmente verosímil que R_1 seguido de R_3 . En consecuencia $p(A \cap B|H)$ es el área de $R_1 \cap R_3$, es decir $p(A|H)$ por $p(B|A, H)$, como queríamos demostrar.

Para nosotros, la probabilidad $p(E|H)$ de un suceso E en las condiciones H es una medida, sobre la escala $[0, 1]$ de la verosimilitud que el decisor concede al suceso E en la situación descrita por H , esto es, una medida del *grado de creencia* en la ocurrencia de E que la información contenida en H le sugiere al decisor.

Hay dos puntos importantes que deben subrayarse. En primer lugar, la probabilidad asignada a un suceso es *siempre* condicional a la información que se posee sobre él; no existen probabilidades «absolutas». En segundo lugar, estamos intentando cuantificar mediante probabilidades la información que *una persona determinada* posee sobre los sucesos inciertos que afectan a las consecuencias de sus decisiones: no existen probabilidades «objetivas». Así, por ejemplo, si denotamos por E el suceso «se lanza un dado y se obtiene un cinco», podríamos escribir $p(E|H) = 1/6$ donde H significa «el decisor juzga que el dado está perfectamente construido»; limitarnos a escribir $p(E) = 1/6$ y pretender que esta fuese una probabilidad «absoluta» y «objetiva» no sería correcto: es posible que el decisor tenga motivos para sospechar que el dado está cargado, en cuyo caso no sería razonable que asignase esa probabilidad.

Análogamente, si V es el suceso «María dará luz a un hijo varón», podríamos escribir $p(V|F, H) = 0,51$ donde F significa «el 51 % de los nacimientos en España producen varones» y H «no hay motivo para suponer que María tenga una tendencia a tener hijos varones distinta de cualquier otra mujer española». Sin embargo, escribir $p(V|F) = 0,51$ y pretender que esta fuese una probabilidad «objetiva» no sería correcto; el árbol genealógico de María podría sugerir una tendencia en su familia a tener más hijos varones que la media de la población.

En aquellas situaciones en las que no existen simetrías aparentes ni datos históricos sobre frecuencias relativas puede parecer más difícil obtener una medida numérica del grado de creencia en un suceso. Sin embargo, no se empieza una escalada si parece *probable* que vaya a haber tormenta ni se realiza una inversión si no parece *probable* que sea rentable; de nuevo, nos vamos continuamente obligados a asignar probabilidades, aunque sea inconscientemente, a todo tipo de sucesos para poder actuar, para poder sobrevivir. Incluso sucesos tan pintorescos como la existencia del monstruo del lago Ness pueden ser objeto de análisis, como demuestra el hecho de que una famosa compañía de seguros aceptase una póliza diseñada para cubrir el riesgo de su aparición (Borch, 1975).

Obviamente, con la definición adoptada, la probabilidad de que un punto aleatorio se sitúe en una región $R(x)$ del cuadrado unidad de área x es precisa-

mente x . En consecuencia, una forma alternativa de describir la situación representada por la opción

$$l(x) = \{c^*|R(x), c_*|\bar{R}(x)\}, \quad (1)$$

que permite obtener c^* con probabilidad x y c_* con la probabilidad complementaria $1 - x$, consiste en escribir simplemente $\{c^*|x, c_*|1 - x\}$. En general, con la notación

$$l = \{c_1|p_1, c_2|p_2, \dots, c_k|p_k\}$$

describiremos una opción que permite obtener la consecuencia c_1 con probabilidad p_1 , c_2 con probabilidad p_2 , ..., c_k con probabilidad p_k .

Ejemplo 2.4.1. Temperaturas

Considérense los sucesos siguientes referidos a la temperatura t en grados centígrados que hará mañana en Valencia a mediodía.

- $A_1 = t < 16$
- $A_2 = 16 \leq t < 19$
- $A_3 = 19 \leq t < 22$
- $A_4 = t \geq 22$

Asígnales una distribución de probabilidad

Claramente, no existe una solución única. El resultado depende de la información del decisor y de la época en que se asigne la distribución de probabilidad. Una distribución razonable para un ciudadano medio, referida a principios de octubre puede ser

$$p(A_1) = 0,05, \quad p(A_2) = 0,15, \quad p(A_3) = 0,60, \quad p(A_4) = 0,20$$

lo que significaría, por ejemplo, que se ha considerado cuatro veces más probable el suceso $t < 22$ que el suceso $t \geq 22$, y también cuatro veces más probable el suceso $t \geq 19$ que el $t < 19$.

2.5. Maximización de la utilidad esperada

Los postulados de coherencia también permiten definir formalmente una medida de las preferencias del decisor entre las consecuencias. En efecto, definiremos la *utilidad* de una consecuencia c como un número $u(c)$ en la escala $[0, 1]$ que mide la deseabilidad relativa de la consecuencia c . En esta escala, como veremos a continuación, la utilidad de la consecuencia menos preferida c_* será $u(c_*) = 0$ y la de la más preferida será $u(c^*) = 1$.

Como en el caso de las probabilidades, necesitamos un elemento de referencia para poder medir. En aquel caso utilizábamos las regiones del cuadrado unidad. Aquí compararemos la deseabilidad de una consecuencia con la de una opción del tipo $\{c^*|x, c_*|1 - x\}$, que permite obtener la consecuencia c^* con probabilidad x , o la c_* con probabilidad $1 - x$.

DEFINICIÓN 2.5.1. La utilidad de una consecuencia c , que denotaremos por $u(c)$ es la probabilidad x que debe asignarse a la mejor consecuencia c^* para que la consecuencia c sea igualmente deseable que la opción $\{c^*|x, c_*|1 - x\}$. De esta forma, para toda consecuencia c ,

$$c \sim \{c^*|u(c), c_*|1 - u(c)\}$$

En virtud de los postulados de comparabilidad, transitividad y sustitución, siempre existirá un número $u(c)$ en $[0, 1]$ que cumpla esa condición puesto que $c^* \sim \{c^*|1, c_*|0\}$, $c_* \sim \{c^*|0, c_*|1\}$ y $c_* \leq c \leq c^*$. Además, la utilidad de c queda así bien definida salvo una transformación lineal que, como veremos más adelante, no afecta a la elección de la decisión óptima.

En efecto, si las consecuencias de referencia c^* y c_* se sustituyen, por ejemplo, por otras consecuencias $c_0 < c_*$ y $c_1 > c^*$, podemos encontrar dos números p_0 y p_1 , las utilidades de c_* y c^* en la nueva escala, tales que $p_0 < p_1$,

$$\begin{aligned} c^* &\sim \{c_1|p_1, c_0|1 - p_1\} \\ c_* &\sim \{c_1|p_0, c_0|1 - p_0\} \end{aligned} \quad (1)$$

Si $u(c)$ es la utilidad de la consecuencia c en la escala primitiva, tenemos por definición que

$$c \sim \{c^*|u(c), c_*|1 - u(c)\} \quad (2)$$

y sustituyendo (1) en (2)

$$c \sim \{c_1|u'(c), c_0|1 - u'(c)\} \quad (3)$$

donde

$$u'(c) = (p_1 - p_0)u(c) + p_0 \quad (4)$$

En consecuencia, la utilidad de c en la nueva escala, $u'(c)$ es una simple transformación lineal de la utilidad original $u(c)$; en particular, $u'(c^*) = p_1$, $u'(c_*) = p_0$. Naturalmente, las utilidades asociadas a los extremos de la nueva escala son, de nuevo, cero y uno: $u'(c_0) = 0$, $u'(c_1) = 1$.

Utilizando los principios de coherencia postulados en la Sección 2.3 hemos podido asignar a los sucesos inciertos unos números (sus probabilidades) que miden la verosimilitud que les asigna el decisor en el momento, y en las condiciones, en que toma la decisión. Análogamente, hemos asignado a las consecuencias otro conjunto de números (sus utilidades) que miden las preferencias del decisor entre ellas. El paso final consiste en asignar un número a cada una de las decisiones posibles de forma que la mejor decisión sea aquella a la que se asigna el número más alto. Esto puede hacerse sin invocar nuevos principios ni hacer nuevas asignaciones numéricas. En efecto, tomar la decisión d_i es aceptar la opción

$$d_i = \{c_{i1}|\theta_{i1}, c_{i2}|\theta_{i2}, \dots, c_{im_i}|\theta_{im_i}\} \quad (5)$$

donde $\Theta_i = \{\theta_{i1}, \theta_{i2}, \dots, \theta_{im_i}\}$ es el conjunto de sucesos inciertos mutuamente excluyentes que pueden afectar a las consecuencias de tomar la decisión d_i y c_{ij} es la consecuencia de haber tomado la decisión d_i cuando sucede θ_{ij} . De acuerdo con las Definiciones 2.41 y 2.5.1 el decisor puede asignar, para cada θ_{ij} la probabilidad $p(\theta_{ij}|d_i, H)$ de que suceda θ_{ij} cuando se elige d_i en las condiciones H , y la utilidad $u(c_{ij})$ de la consecuencia a que esto da lugar.

TEOREMA 2.5.1. (Criterio de decisión Bayes) *Considérese el problema de decisión definido por $D = \{d_1, d_2, \dots, d_k\}$ donde*

$$d_i = \{c_{i1}|\theta_{i1}, c_{i2}|\theta_{i2}, \dots, c_{im_i}|\theta_{im_i}\}$$

Sea $p(\theta_{ij}|d_i, H)$ la probabilidad de que suceda θ_{ij} si se elige d_i en las condiciones H y sea $u(c_{ij})$ la utilidad de la consecuencia a que ello da lugar. Entonces, la utilidad esperada de la decisión d_i es

$$u^*(d_i) = \sum_{j=1}^{m_i} u(c_{ij})p(\theta_{ij}|d_i, H) \quad (6)$$

y la decisión óptima es aquella con máxima utilidad esperada

Demostración

En efecto, por definición de utilidad,

$$c_{ij} \sim \{c^*|u(c_{ij}), c_*|[1 - u(c_{ij})]\} \quad (7)$$

y por definición de probabilidad,

$$d_i \sim \{c_{i1}|p(\theta_{i1}|d_i, H), c_{i2}|p(\theta_{i2}|d_i, H), \dots, c_{im_i}|p(\theta_{im_i}|d_i, H)\} \quad (8)$$

Introduciendo (7) y (8) en virtud del postulado de sustitución y del Teorema 2.4.1,

$$\begin{aligned} d_i &\sim \{c^*|u(c_{i1})p(\theta_{i1}|d_i, H), c_*|[1 - u(c_{i1})]p(\theta_{i1}|d_i, H), \\ &c^*|u(c_{i2})p(\theta_{i2}|d_i, H), c_*|[1 - u(c_{i2})]p(\theta_{i2}|d_i, H), \\ &c^*|u(c_{im_i})p(\theta_{im_i}|d_i, H), c_*|[1 - u(c_{im_i})]p(\theta_{im_i}|d_i, H)\} \end{aligned}$$

y, por tanto,

$$d_i \sim \{c^*|u^*(d_i), c_*|[1 - u^*(d_i)]\} \quad (9)$$

donde $u^*(d_i)$ viene dado por (6). Ahora bien, en la forma (9) todas las decisiones son inmediatamente comparables y, en virtud de la transitividad, la más preferible es aquella que da más probabilidad a la consecuencia óptima c^* , es decir aquella que maximiza $u^*(d_i)$ como queríamos demostrar.

Si la función de utilidad $u_1(c)$ se sustituye por una transformación lineal cuya $u_2(c) = au_1(c) + b$, la nueva utilidad esperada es claramente

$$u_2^*(d) = au_1^*(d) + b$$

Así pues, si $a > 0$, la decisión que maximiza $u_1^*(d)$ también maximiza $u_2^*(d)$ y por tanto, como habíamos anticipado, la elección de las consecuencias de referencia c_* y c^* no afecta a la determinación de la decisión óptima.

Aunque la función de utilidad ha sido definida en la escala $[0, 1]$, resulta a veces más natural medir la deseabilidad de las consecuencias en una escala distinta (tiempo, dinero, años de vida, ...) Siempre es posible, sin embargo, reducir las utilidades así obtenidas a la escala $[0, 1]$ mediante la transformación

$$u'(c) = [u(c) - u(c_*)]/[u(c^*) - u(c_*)] \quad (10)$$

donde c^* y c_* son, respectivamente, la mejor y la peor de las consecuencias consideradas. Sin embargo, puesto que (10) es una transformación lineal, tal reducción no es necesaria, según hemos visto, para determinar la decisión óptima. El uso de la escala $[0, 1]$ tiene, no obstante, la ventaja de permitir una interpretación probabilista de las utilidades.

Ejemplo 2.5.1. Opciones económicas alternativas (cont.)

Considérese de nuevo la elección entre las opciones l_1, l_2, l_3 y l_4 consideradas en el Ejemplo 2.3.1. Asignar probabilidades a los sucesos inciertos y

determinar la utilidad esperada de cada una de ellas suponiendo la utilidad proporcional al dinero.

Si, por ejemplo, se considera $p(A_1|H) = 0,3$ y $p(A_2|H) = 0,6$

$$u^*(l_1) = 1\,000 \times 0,3 - 100 \times (1 - 0,3) = -260$$

$$u^*(l_2) = 500 \times 0,6 - 100 \times (1 - 0,6) = 260$$

$$u^*(l_3) = 200 \times 0,3 \times 0,6 + 100 \times 0,3 \times 0,4 - 150 \times 0,7 = -57$$

$$u^*(l_4) = 400 \times 0,3 \times 0,6 + 300 \times 0,7 \times 0,6 - 350 \times 0,4 = 58$$

Estas utilidades esperadas implican las relaciones $l_1 < l_3 < l_4 < l_2$, que son las mencionadas en el Ejemplo 2.3.1. El lector puede comprobar sin embargo que esta estructura de preferencias entre las opciones depende *crucialmente* de la probabilidades asignadas a los sucesos A_1 y A_2 en las condiciones H en que se efectúa la elección.

2.6. Otros criterios de decisión

Comentaremos a continuación el alcance y las limitaciones del criterio de decisión establecido, el *criterio Bayes* o de maximización de la utilidad esperada, y lo compararemos con otros criterios de decisión propuestos en la literatura.

Probablemente, la mayor limitación de la teoría esbozada radica en el hecho de que su uso está restringido a un solo decisor y no es inmediatamente aplicable, supuesto que pueda serlo, a situaciones que involucran a dos o más decisores. Así, nos hemos referido continuamente a un único decisor, dispuesto a expresar sus opiniones y sus preferencias y a ser consistente con ellas, y hemos concluido que tal persona debe elegir la decisión que maximice su utilidad esperada; sin embargo, no hay nada en el desarrollo precedente que obligue a dos decisores distintos a ponerse de acuerdo en las probabilidades asignadas a los sucesos inciertos o en las utilidades dadas a las posibles consecuencias.

A menudo, una decisión debe ser tomada por un organismo colegiado, por un comité, por una asamblea. En este caso, puede suponerse frecuentemente que todos los componentes tienen los mismos objetivos, y por tanto las mismas utilidades, pero que difieren en su apreciación de la realidad, esto es en sus probabilidades. En otros casos sin embargo, las preferencias de los decisores son claramente contrapuestas entre sí; sus utilidades (y tal vez sus probabilidades) son diferentes y nos encontramos en consecuencia en una típica situación de conflicto. No existe una teoría axiomática que resuelva el problema planteado por las situaciones de comité o de conflicto de forma comparable a la solución que la teoría expuesta ofrece para el problema de decisión *unipersonal*. La consecución de tal teoría es una de las mayores necesidades de nuestra época.

La fuerza del criterio de maximización de la utilidad esperada reside esencialmente en su fundamento axiomático. Es el único criterio de decisión compatible con los principios de coherencia expuestos en la Sección 2.3. Cualquier otro criterio es equivalente a la maximización de una utilidad esperada, o es incoherente. En el resto de esta sección comentaremos brevemente algunas críticas hechas al principio de la utilidad esperada y ejemplificaremos la inconsistencia de otros criterios de decisión.

Una de las objeciones frecuentes al principio de maximización de la utilidad esperada consiste en afirmar que no tiene en cuenta el riesgo. Por ejemplo, una persona puede afirmar que prefiere 100 pesetas seguras a 101 con probabilidad p y nada con *probabilidad* $1 - p$ por próximo a uno que sea el valor de p . ¿Por qué arriesgarse a perder 100 por la posibilidad de ganar un poco más? Es fácil ver que esta afirmación, que obviamente viola el principio de maximización de la utilidad esperada, es incoherente.

En efecto, si

$$100 > \{101|p, 0|(1-p)\}$$

entonces, presumiblemente,

$$101 > \{102|p, 0|(1-p)\}$$

y por lo tanto

$$100 > \{102|p^2, 0|(1-p^2)\}$$

y prosiguiendo de esta forma

$$100 > \{1000|p^{999}, 0|1-p^{999}\}$$

Sin embargo, para un p tal que p^{999} sea suficientemente grande, cualquiera preferiría 1000 con esa probabilidad a 100 seguras, lo que, según hemos visto, contradice la afirmación original.

Otra forma de decir que el principio de la utilidad esperada no tiene en cuenta el riesgo consiste en sostener que, en cualquier situación de posible inversión, maximizar la utilidad esperada implicaría invertir todo el capital en el tipo de acciones que se espere sea más rentable, en contra del principio empíricamente establecido de la diversificación. Quien utiliza este argumento está confundiendo las cantidades de dinero con su utilidad. La mayoría de nosotros tenemos una función de utilidad cóncava para el dinero, con una utilidad marginal decreciente para cantidades adicionales. Puede comprobarse (Lindley, 1971 b, cap. 5) que la decisión óptima que se deduce de maximizar la utilidad esperada con una función cóncava para la utilidad del dinero es generalmente la de diversificar la inversión (ver Sección 7.3).

A juzgar por los textos clásicos de teoría de la decisión, el criterio *minimax* podría ser una alternativa razonable al criterio de maximización de la utilidad esperada. En una de sus versiones, el criterio *minimax* consiste en tomar la decisión que *maximiza la utilidad garantizada*: se determina lo peor que puede pasar en cada caso, y se elige aquella decisión para la que resulta menos malo. Más técnicamente, la decisión óptima según este criterio es la que maximiza la utilidad mínima a que puede dar lugar.

Así, en el problema de decisión definido por la Tabla

	θ_1	θ_2	mín
d_1	1	0	0
d_2	0,3	0,3	0,3 máx

la decisión óptima según el criterio *minimax* es d_2 , puesto que d_2 maximiza la utilidad garantizada. En consecuencia, según este criterio, $d_2 > d_1$. De forma análoga, en el problema de decisión definido por la Tabla

	θ_1	θ_2	mín
d_2	0,3	0,3	0,3 máx
d_3	0	1	0

encontraremos $d_2 > d_3$. Consideremos una nueva decisión d_4 que consiste en elegir d_1 si sale cara y d_3 si sale cruz. Considerando ahora todas estas decisiones juntas tendríamos

	θ_1	θ_2	mín
d_1	1	0	0
d_2	0,3	0,3	0,3
d_3	0	1	0
d_4	0,5	0,5	0,5 máx

con lo que $d_4 > d_2$. Tenemos pues una situación en la que se prefiere d_2 tanto a d_1 como a d_3 , pero en la que una elección d_1 y d_3 basada en lanzar una moneda al aire es preferible a d_2 . Esto es incoherencia.

En otra versión del método *minimax* se escoge aquella decisión que hace mínima la mayor *pérdida de oportunidad* posible, entendiendo por tal la diferencia entre la utilidad conseguida y la que hubiese podido conseguir.

Consideremos el problema de decisión definido por las tablas siguientes (utilidades a la izquierda, pérdidas de oportunidad a la derecha)

	θ_1	θ_2	máx
d_1	0,8	0,0	0,4 mín
d_2	0,2	0,4	0,6

Claramente, según el criterio *minimax*, debe preferirse d_1 a d_2 . Consideremos ahora una tercera decisión posible, d_3 , de forma que las nuevas tablas de utilidades y pérdidas de oportunidad resultan ser

	θ_1	θ_2	máx
d_1	0,8	0,0	0,7
d_2	0,2	0,4	0,6 mín
d_3	0,1	0,7	0,7

En este caso, el criterio *minimax* sugiere la elección de d_2 . Resulta así que la consecuencia de incorporar una nueva decisión d_3 al conjunto de posibles alternativas es nada menos que sustituir $d_1 > d_2$ por $d_2 > d_1$. Es como si se prefiriese estudiar Medicina a Ciencias pero, *por el hecho* de pensar también en Filosofía, se concluyese que se preferiría Ciencias a Medicina. Esto es incoherencia.

A nivel intuitivo, la causa de la incoherencia del método *minimax* reside tanto en su excesivo pesimismo como en el hecho de que no utiliza la información de que se dispone sobre la verosimilitud relativa de los sucesos. Desde un punto de vista técnico, la causa de la incoherencia está en el hecho de que el criterio *minimax* no satisface el postulado de sustitución.

Otro criterio de decisión frecuentemente empleado en la práctica es el *criterio condicional* que consiste en tomar la decisión que resultaría óptima si el suceso más probable tuviese efectivamente lugar. Considérese, por ejemplo, el problema de decisión descrito por la tabla de utilidades

	θ_1	θ_2	θ_3
d_1	0,5	0,5	0,6
d_2	0,4	0,4	0,7

y supongamos que $p(\theta_1) = 0,30$, $p(\theta_2) = 0,25$, $p(\theta_3) = 0,45$. De acuerdo con este criterio, $d_2 > d_1$ puesto que si el suceso más probable (θ_3) tiene lugar, da una utilidad mayor. Sin embargo, observando que las utilidades corres-

pendientes a d_1 y d_2 coinciden, el problema de decisión puede reformularse mediante la tabla

	$\theta_1 \cup \theta_2$	θ_3
d_1	0,5	0,6
d_2	0,4	0,7

donde, obviamente $p(\theta_1 \cup \theta_2) = 0,55$ y $p(\theta_3) = 0,45$. El suceso más probable es ahora $(\theta_1 \cup \theta_2)$ y por tanto $d_1 > d_2$. De nuevo, esto es incoherente. A nivel intuitivo el método descrito, desgraciadamente muy utilizado en la práctica, es menos razonable de lo que puede parecer a primera vista. En efecto, puede suceder que una decisión que no sea óptima para la más probable de las posibilidades sea sustancialmente mejor que las demás en el resto de ellas, y resulte mejor en conjunto.

Ejemplo 2.6.1. Forma de estudio

Un alumno debe decidir si repasar con mucho detalle una de las dos partes del programa o repasar con menos detalle las dos, antes de examinarse. Analizar el problema según los distintos criterios de decisión mencionados, suponiendo que juzga que lo más probable es que el examen contenga mayoritariamente cuestiones sobre la segunda parte.

El espacio de decisiones es

- d_1 = repasar con detalle la primera parte
- d_2 = repasar con detalle la segunda parte
- d_3 = repasar todo el programa con menos detalle

y el de sucesos inciertos puede describirse como

- θ_1 = el examen contiene mayoritariamente temas de la primera parte
- θ_2 = el examen contiene mayoritariamente temas de la segunda parte
- θ_3 = el examen está equilibrado

Una tabla razonable de utilidades sería entonces del tipo (otras tablas parecidas serían igualmente aceptables).

	θ_1	θ_2	θ_3	mín
d_1	0,9	0,2	0,5	0,2
d_2	0,2	0,9	0,5	0,2
d_3	0,6	0,6	0,7	0,6 máx

El criterio minimax recomienda elegir siempre d_3 , independientemente de las verosimilitudes relativas de θ_1 y θ_2 . El criterio de maximización de la utilidad correspondiente al suceso más probable recomienda elegir siempre d_2 , puesto que por hipótesis $p(\theta_2) > p(\theta_3)$ y $p(\theta_2) > p(\theta_1)$. Ambas soluciones son claramente demasiado rígidas.

En efecto, las utilidades esperadas de las tres decisiones son

$$u^*(d_1) = 0,9p + 0,2q + 0,5(1 - p - q) = 0,5 + 0,4p - 0,3q$$

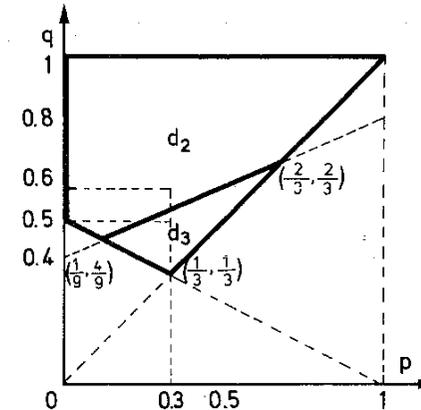
$$u^*(d_2) = 0,2p + 0,9q + 0,5(1 - p - q) = 0,5 - 0,3p + 0,4q$$

$$u^*(d_3) = 0,6p + 0,6q + 0,7(1 - p - q) = 0,7 - 0,1p - 0,1q$$

donde $p = p(\theta_1)$, $q = p(\theta_2)$ y, por hipótesis, $q > p$ y $q > 1 - p - q$, esto es $p > 1 - 2q$.

Es inmediato comprobar que $u^*(d_1) > u^*(d_2)$ si, y solamente si $p > q$; como, por hipótesis $q > p$, d_1 siempre es mejor que d_2 . Técnicamente, d_1 es una solución *inadmisible* para este problema.

Puede comprobarse además que $u^*(d_2) > u^*(d_3)$ si, y solamente si, $0,2p - 0,5q + 0,2 < 0$. En la figura se representan los conjuntos de pares de valores (p, q) para los que la decisión óptima es una u otra.



Por ejemplo, si $p = 1/3$ y $q = 0,5$, la decisión óptima es d_3 , esto es repasar todo el programa, mientras que si $p = 1/3$ pero $q = 0,6$, la decisión óptima es d_2 , es decir repasar con detalle la segunda parte.

2.7. Discusión y referencias

Los resultados fundamentales del argumento Bayesiano pueden sintetizarse diciendo que la consideración de unos principios razonables de comportamiento llevan a la existencia de una medida de probabilidad, de una función de utilidad y al principio de maximización de la utilidad esperada. Sin embargo, los resultados mencionados son tan sólo *teoremas de existencia*. Debe subrayarse que, aunque las definiciones de probabilidad y de utilidad son constructivas, no se trata necesariamente de los mejores métodos para deter-

minarlas. En los Capítulos 3 y 7 discutiremos otros procedimientos para la especificación respectivamente, de probabilidades y de utilidades.

El argumento Bayesiano no es, desde luego, nuevo. En esencia se remonta a los trabajos de Bayes (1763) y Laplace (1812/1912) y siempre ha merecido respeto entre los estadísticos. Sin embargo, en años recientes ha aparecido en la discusión un nuevo elemento: el hecho de que, en cierto sentido, los métodos Bayesianos son inevitables porque se deducen de una axiomática que resulta razonable y frente a la que no se ha sabido presentar, con fuerza parecida, alternativa alguna.

La primera discusión axiomática se debe a Ramsey (1926). Su trabajo, cuya línea hemos seguido en este capítulo, carece de demostraciones rigurosas, como suele suceder con las obras de los matemáticos aplicados británicos de los años veinte. Tales demostraciones fueron proporcionadas mucho más tarde por Savage (1954/1961). Una demostración rigurosa de los resultados Bayesianos en el caso finito es la de Pratt, Raiffa & Schaleifer (1964), basada en un trabajo previo de Anscombe y Aumann (1963); Bernardo y Giron (1980) proporcionan una demostración rigurosa en el caso general. Villegas (1964) y DeGroot (1970) desarrollan por separado las axiomas de la probabilidad y de la utilidad.

Por otra parte, De Finetti (1937) produjo en los años treinta, sin conocer el trabajo de Ramsey, un tipo de argumento muy distinto para la existencia de probabilidades subjetivas. Se basa en considerar un sistema de apuestas y exigir que se hagan de forma que no sea posible perder siempre. El razonamiento tiene el inconveniente de que, al introducir apuestas, se crea cierta confusión con ideas de utilidad. Los artículos originales de Ramsey y de De Finetti han sido reeditados, junto con otros clásicos de la probabilidad subjetiva, como el de Koopman (1940), por Kyburg & Smokler (1964). El concepto de coherencia ha sido brillantemente discutido por Cornfield (1969) en relación con la estadística moderna.

Una línea de investigación en Teoría de la Decisión, relacionada con la idea de coherencia, pero mucho más débil que ella, parte del trabajo de Wald (1950). Sus resultados son más débiles debido a que *supone* la existencia de una función de pérdida, en lugar de *deducir* su existencia a partir de los axiomas.

Las objeciones planteadas a la postura Bayesiana son numerosas, pero ninguna ha llegado al fondo de la cuestión y criticado constructivamente los axiomas. Una excelente discusión sobre los fundamentos de la Estadística es la contenida en Savage *et al.* (1962). Lindley (1971 b, cap 1-4) expone de forma muy intuitiva y con mucho más detalle los resultados comentados en este capítulo.

PROBLEMAS

1. Un juego consiste en elegir al azar una carta de una baraja de 40 cartas. Si esa carta es una espada nos pagan 200 ptas. y si es oro 100 ptas., pero en otro caso hemos de pagar 100 ptas. Para empezar a jugar hay que poner una cantidad inicial de c ptas. Igualando dinero a utilidad ¿Interesará jugar si $c = 30$ ptas.? ¿Qué valor de c haría el juego *equilibrado*? (Se dice que un juego es equilibrado si la ganancia esperada de jugar es igual a la de no jugar.)
2. El beneficio esperado de la comercialización de un nuevo fármaco es de 2 millones de pesetas si su puesta en venta es anterior a la de un preparado similar ofrecido por la competencia, pero se perdería un millón si la competencia se adelanta. Determinar la decisión óptima en función de la probabilidad p de conseguir adelantarse a la competencia.
3. En una prueba clínica se intenta comparar dos cremas diferentes que previenen la reaparición de una determinada alergia de la piel. Después de probadas con 150 pacientes, obtendremos la siguiente tabla de *probabilidades*:

	No reaparece	Reaparece
crema 1	0,7	0,3
crema 2	0,6	0,4

Como la crema 1 produce determinados efectos secundarios, esto nos influye en las *utilidades* que resultan ser:

	No reaparece	Reaparece
crema d_1	0,9	0
crema d_2	1	0,1

Aparece un nuevo paciente con ese tipo de alergia. Determinar la crema que debe recetársele.

4. La esperanza de vida de un determinado paciente de cáncer es 4 años. Si se opera, y la operación sale bien, su esperanza de vida aumenta a 10 años, pero tiene una probabilidad p de morir en la operación y una probabilidad 0,5 de quedarse como estaba a pesar de operarse. ¿Qué valores de p hacen deseable la operación?
5. Ante un caso de ictericia que puede haber sido causado por una hepatitis o por un cáncer de páncreas, el médico debe elegir entre un tratamiento clínico y uno quirúrgico. La esperanza de vida del paciente en el caso de operar es de 10 años si tiene realmente cáncer; si tiene hepatitis y es operado, tiene una probabilidad 0,2 de morir como consecuencia de la operación y una esperanza de 15 años de vida si sobrevive a ella. La esperanza de vida del paciente en el caso de no operar es de 2 años si tiene cáncer y de 16 si tiene hepatitis. Sea p la probabilidad que el médico asigna a que la causa de la ictericia sea hepatitis. ¿Qué valores de p hacen deseable la operación?
6. Un individuo tiene un cierto lote de productos. Si lo vende gana 400 pesetas y si no lo puede vender pierde 300. Puede anunciar o no. Anunciar le cuesta 200 ptas. Si pone el anuncio la probabilidad de vender es $4/5$. Determinar la decisión óptima.

en función de la probabilidad p de vender si no anuncia. Supóngase la utilidad proporcional al dinero.

7. Se dispone de dos urnas, que denotaremos por I y II. La urna I contiene 3 bolas rojas y 2 blancas. La urna II contiene 5 rojas y 4 blancas. El juego consiste en sacar un máximo de dos bolas. Por cada bola extraída se gana 100 ptas, si es roja, y se pierde 50 ptas, si es blanca. Por la primera extracción no hay que pagar nada. Por la segunda extracción hay que pagar 50 ptas, si se trata de la urna I y 25 si se trata de la urna II. Determinar la estrategia óptima.
8. Durante los días previos a determinadas elecciones al Parlamento británico, se ofrecían en distintas casas de apuestas londinenses apuestas en proporción de 4 a 7 contra que ganasen los laboristas (una persona que apostase a favor de los laboristas *ganaría* 7 libras por cada 4 que hubiese invertido si los laboristas ganasen las elecciones) y en proporción de 5 a 4 contra que ganasen los conservadores (una persona que apostase a favor de los conservadores *ganaría* 4 libras por cada 5 que hubiese invertido si los conservadores ganasen las elecciones). Suponiendo la utilidad proporcional al dinero, determinar la estrategia de juego óptima en función de la probabilidad personal p que se asigne a una victoria laborista.
9. Una compañía que fabrica cierta maquinaria sofisticada, debe decidir su plan de producción mensual que puede consistir en fabricar 1, 2 o 3 máquinas. Sabiendo que la demanda puede ser de 0, 1, 2, 3, 4 máquinas al mes y que la probabilidad de que la demanda sea 2 es 0,4 y el resto de las posibilidades son igualmente probables, encontrar un plan de producción óptimo sabiendo que hay un beneficio de 7 por unidad vendida, una pérdida de 4 por unidad no satisfecha y una pérdida de 1 por unidad almacenada. Considerar la utilidad proporcional al beneficio.
10. Un médico cree que su paciente tiene una y sólo una de las enfermedades $\theta_1, \theta_2, \theta_3, \theta_4$, y que sus probabilidades respectivas son $p(\theta_1) = 0,2$, $p(\theta_2) = 0,5$, $p(\theta_3) = 0,2$, $p(\theta_4) = 0,1$. El médico dispone de tres tratamientos alternativos t_1, t_2 y t_3 cuya efectividad viene reflejada en la siguiente tabla.

	θ_1	θ_2	θ_3	θ_4
t_1	1	0	1	1
t_2	0	1	0	1
t_3	0	1	1	0

donde un uno es la fila i y columna j significa que el tratamiento t_i es efectivo contra la enfermedad θ_j y un cero significa que no lo es. Igualando utilidad a efectividad, determinar el tratamiento óptimo.

Considérese que si un tratamiento no es efectivo puede procederse a otro y determinarse la sucesión óptima de tratamientos, suponiendo que sus costes respectivos son $c(t_1) = 0,1$, $c(t_2) = 0,2$, $c(t_3) = 0,3$ en unidades de utilidad.

Medida de probabilidad

En la Sección 2.4 se introdujo el concepto de probabilidad como *grado de creencia* y se dedujeron algunas de sus propiedades. En este capítulo se estudian las consecuencias matemáticas que se deducen de estas propiedades con objeto de poder describir adecuadamente los fenómenos aleatorios y de saber expresar convenientemente nuestra información sobre su verosimilitud.

Concretamente, se introduce el concepto de *espacio probabilístico* con objeto de especificar el conjunto de sucesos para los que la medida de probabilidad está definida. Se comentan las propiedades fundamentales de la probabilidad y los teoremas básicos a que da lugar. Se estudian los conceptos de *independencia* e *intercambiabilidad* y se enuncian y comentan los Teoremas de Bayes y de la *probabilidad total*. Finalmente, se estudian algunos métodos para la especificación de probabilidades.

En el capítulo anterior, definimos la probabilidad de un suceso A en las condiciones H , que escribíamos $p(A|H)$, como una medida del grado de creencia en A que sugiere al decisor la información contenida en H . Sin embargo, la palabra *suceso* fue utilizada en sentido coloquial, sin precisar el espacio sobre el que la medida de probabilidad debía ser definida. En la próxima sección especificaremos este extremo.

3.1. Espacios probabilísticos

Supondremos que en la descripción de cualquier situación incierta existe un *conjunto de referencia* Ω que contiene como subconjuntos a todos los sucesos en cuya probabilidad estamos interesados. En un problema de diag-

nosis médica, el conjunto de referencia puede ser el de las enfermedades compatibles con los síntomas observados. Si tratamos en cambio de medir la altura de una persona, el conjunto de referencia será la recta real positiva o, mejor, un subconjunto suyo.

En general, no estaremos interesados en asignar probabilidades a todos los subconjuntos de Ω . ¿Qué utilidad tendría, por ejemplo, determinar la probabilidad de que la altura que pretendemos medir sea un número entero impar? Por otra parte, es razonable suponer que si nos interesa la probabilidad de determinados sucesos, nos interesará igualmente la de sus uniones, sus intersecciones y sus complementos. Así, si nos preguntamos la probabilidad de que una persona mida más de 1,50 mts, y también la de que mida menos de 1,80, posiblemente nos interese la probabilidad de que su altura se sitúe entre 1,50 y 1,80 mts. Esto nos lleva a exigir una determinada estructura algebraica para el conjunto de sucesos sobre el que la medida de probabilidad debe estar definida (*).

DEFINICIÓN 3.1.1. Una familia de conjuntos está dotada de una estructura de álgebra si el conjunto vacío pertenece a ella y son leyes de composición interna la unión y la complementación

Es fácil comprobar que nuestra definición implica, en particular, que el conjunto de referencia pertenece al álgebra y que la intersección también es una ley de composición interna. Así, si A y B son dos subconjuntos del conjunto de referencia Ω que pertenecen al álgebra de conjuntos en que estamos interesados, \bar{A} , \bar{B} , $A \cup B$, $A \cap B$, $A \cup \bar{B}$, $A \cap \bar{B}$, etc., también pertenecerán a ella. El álgebra engendrada por un conjunto de sucesos interesantes es la familia de todos los conjuntos que pueden formarse a partir de ellos mediante las operaciones de unión, intersección y complemento. Así, partiendo de un único suceso interesante A , se obtiene un álgebra de $2^2 = 4$ elementos:

$$\Sigma = \{\Omega, \phi, A, \bar{A}\}$$

donde $\Omega = A \cup \bar{A}$ y $\phi = A \cap \bar{A}$. Partiendo de dos conjuntos distintos se puede engendrar un álgebra de $2^4 = 16$ elementos; en general, partiendo de k conjuntos interesantes distintos se engendra un álgebra con 2^k elementos como máximo.

Nuestra descripción de una situación incierta cuyo conjunto de referencia es Ω quedará completa mediante la especificación del álgebra de subcon-

(*) Puede demostrarse, además, que admitir como sucesos a todos los subconjuntos de un conjunto de referencia puede plantear graves problemas para la definición de sus probabilidades cuando el conjunto de referencia tiene infinitos elementos.

juntos de Ω engendrada por aquellos sucesos en cuya probabilidad estamos interesados y la especificación de sus correspondientes probabilidades.

DEFINICIÓN 3.1.2. Un espacio probabilístico es un conjunto Ω dotado de un álgebra Σ de subconjuntos suyos y de una medida de probabilidad P definida sobre ellos, lo representaremos mediante el triplete (Ω, Σ, P) .

Naturalmente, en virtud de nuestra definición de probabilidad, la medida de probabilidad P estará únicamente definida en las condiciones H que describen la situación incierta; de esta forma, tendremos definidas las probabilidades $p(A|H)$ para todo suceso A que pertenezca al álgebra Σ y, además, tales probabilidades verificarán las propiedades descritas en el Teorema 2.4.1.

Ejemplo 3.1.1. Número de hemáties en la sangre

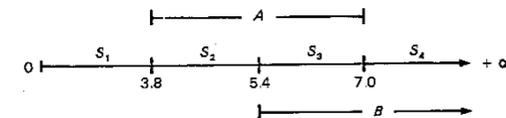
El número medio de hemáties de un varón adulto es de 5,4 millones por milímetro cúbico en sangre venosa periférica, pero se considera «normal» cualquier cantidad comprendida entre 3,8 y 7,0 millones/mm³. Suponiendo que se esté interesado en la probabilidad de que el número de hemáties esté dentro de límites normales y también en la probabilidad de que esté por encima del valor medio, construir el espacio probabilístico adecuado.

El conjunto de referencia Ω es el conjunto de los números reales positivos, puesto que el número de hemáties no puede ser negativo, y no se ha determinado una cota superior. Por hipótesis, nos interesa la probabilidad de los sucesos

$$A = \{x; 3,8 < x < 7,0\}$$

$$B = \{x; x > 5,4\}$$

donde x es el número de hemáties en millones/mm³. En virtud de la Definición 3.1.1, si A y B son elementos del álgebra de sucesos inciertos, también lo serán \bar{A} , \bar{B} y todas las uniones e intersecciones que pueden hacerse con ellos.



El álgebra de sucesos puede obtenerse como el conjunto de las partes del conjunto $\{S_1, S_2, S_3, S_4\}$ donde

$$S_1 = \bar{A} \cap \bar{B} = \{x; 0 \leq x \leq 3,8\}$$

$$S_2 = A \cap \bar{B} = \{x; 3,8 < x \leq 5,4\}$$

$$S_3 = A \cap B = \{x; 5,4 < x \leq 7,0\}$$

$$S_4 = \bar{A} \cap B = \{x; x > 7,0\}$$

y contendrá, por lo tanto $2^4 = 16$ elementos. Puesto que los conjuntos $\{S_1, S_2, S_3, S_4\}$ forman una partición de $\Omega = [0, +\infty[$, bastará asignarles probabilidades a estos cuatro; las probabilidades de los demás se pueden calcular entonces inmediatamente teniendo en cuenta que son disjuntos y que, por lo tanto, en virtud del Teorema 2.4.1, la probabilidad de la unión de dos o más de ellos es la suma de sus correspondientes probabilidades. En el problema que nos ocupa, los datos acumulados en los bancos de sangre permiten asegurar que, para un paciente «medio»,

$$p(S_1) = p(S_4) = 0,025$$

$$p(S_2) = p(S_3) = 0,475$$

Todos los demás sucesos del álgebra, excepto el conjunto vacío, de probabilidad igual a cero, serán uniones de los S_i , y sus probabilidades se pueden determinar por lo tanto con la fórmula $p(U S_i) = \sum p(S_i)$. En particular,

$$p(A) = p(S_2 \cup S_3) = p(S_2) + p(S_3) = 0,95$$

$$p(B) = p(S_3 \cup S_4) = p(S_3) + p(S_4) = 0,5$$

3.2. Teoremas básicos

De nuestra definición de probabilidad, en términos del área correspondiente a un suceso del experimento auxiliar igualmente verosímil, pudimos deducir (Teorema 2.4.1) unas propiedades elementales que permiten relacionar entre sí las probabilidades de distintos sucesos. Como se recordará, el Teorema 2.4.1 nos garantiza que para todo par de sucesos A, H y cualesquiera que sean las condiciones H en que se asignan probabilidades,

$$P2. \quad 0 \leq p(A|H) \leq 1 \quad \text{y} \quad p(H|H) = 1$$

$$P2. \quad \text{Si dado } H, A \cap B = \phi, \text{ entonces } p(A \cup B|H) = p(A|H) + p(B|H)$$

$$P3. \quad p(A \cap B|H) = p(A|H) p(B|A, H)$$

Debe subrayarse que si se conocen las probabilidades correspondientes a un conjunto de sucesos, las propiedades P1, P2 y P3 permiten determinar las probabilidades de todos los sucesos pertenecientes al álgebra que engendran

Las propiedades P1, P2 y P3, que nosotros hemos obtenido como una consecuencia matemática de nuestra definición de probabilidad y de los postulados de coherencia, pueden ser propuestas como *postulados*, para una definición axiomática de la probabilidad. En tal caso, tales axiomas o postulados suelen ser justificados en términos frecuentistas, considerando que la probabilidad de un suceso debe ser un número cercano a su frecuencia relativa.

Para verlo, consideremos la frecuencia relativa con que tiene lugar un suceso A en las condiciones H . Por ejemplo, H podría describir el lanzamiento de una chincheta al aire y A el suceso de que la chincheta repose finalmente con la punta hacia arriba. Para un frecuentista, $p(A|H)$ sería entonces el número alrededor del cual oscila el cociente m/n , donde m es el número de veces que la chincheta queda con la punta hacia arriba y n el número total de lanzamientos. Claramente, una frecuencia relativa es un número entre 0 y 1 y, además, la frecuencia relativa de un suceso que necesariamente ha de tener lugar es uno: esto explica la primera propiedad, P1

Por otra parte, consideremos dos sucesos A, B mutuamente excluyentes en las condiciones H . Si el suceso A tiene lugar m_1 veces entre las n , el suceso B m_2 entre las n , y ambos sucesos son disjuntos, el suceso $A \cup B$ tiene lugar $m_1 + m_2$ veces y por tanto su frecuencia relativa es la suma de las frecuencias relativas de A y de B . Esto explica P2

Finalmente, consideremos dos sucesos cualesquiera A y B . Si, en las condiciones H , el suceso A tiene lugar m veces de un total de n y en r de esas veces también sucede B , entonces r/n es la frecuencia relativa de $A \cap B$ dado H , m/n la de A en las mismas condiciones y r/m la frecuencia relativa de B dados A y H . Pero, trivialmente, $r/n = (m/n)(r/m)$; esto explica P3.

Desde un punto de vista matemático, no hay objeción alguna a una definición axiomática de la probabilidad. Sin embargo, la justificación de las propiedades P1, P2 y P3 en términos de frecuencias relativas no autoriza a suponer que estas propiedades permanezcan válidas en términos de grados de creencia, ni a justificar el uso de probabilidades en un problema de decisión. La importancia del Teorema 2.4.1 radica en demostrar que los grados de creencia, cuya introducción es necesaria según vimos para resolver un problema de decisión de forma coherente, tienen las mismas propiedades matemáticas que las frecuencias relativas y se comportan por lo tanto como las probabilidades clásicas.

Las propiedades P2 y P3, a las que se llaman respectivamente leyes *aditiva* y *multiplicativa* de la probabilidad, se pueden generalizar sin dificultad a un número finito de sucesos por un proceso de inducción. Así,

TEOREMA 3.2.1. Si A_1, A_2, \dots, A_k son sucesos mutuamente excluyentes en las condiciones H ,

$$p\left(\bigcup_{i=1}^k A_i|H\right) = \sum_{i=1}^k p(A_i|H)$$

La demostración del Teorema 3.2.1 a partir de la propiedad P2 es trivial. Sin embargo, no es posible *deducir* que este resultado sea válido en el caso infinito, de forma que

si $\{A_i, i = 1, 2, \dots\}$ son sucesos mutuamente excluyentes en las condiciones H , no es posible deducir que

$$p\left(\bigcup_{i=1}^{\infty} A_i|H\right) = \sum_{i=1}^{\infty} p(A_i|H) \quad (1)$$

En la línea de la mayor parte de los textos de probabilidad, postularemos, sin embargo, que (1) es cierto. La razón de este nuevo postulado, llamado de σ -aditividad, es exclusivamente de tipo matemático, y aparecerá en el capítulo siguiente.

TEOREMA 3.2.2. Para toda colección finita de sucesos A_1, A_2, \dots, A_k ,

$$p\left(\bigcap_{i=1}^k A_i|H\right) = p(A_1|H) p(A_2|A_1, H) \dots p(A_k|A_1, A_2, \dots, A_{k-1}, H)$$

De nuevo la demostración es inmediata por inducción a partir de la propiedad P3. Si se acepta el postulado (1) de σ -aditividad, el teorema 3.2.2 puede extenderse también al caso infinito.

De las propiedades P1, P2 y P3 puede deducirse inmediatamente otras propiedades de la medida de probabilidad. Así por ejemplo

TEOREMA 3.2.3. Para todo par de sucesos A y B , y para cualesquiera condiciones H ,

- (i) $p(\Phi|H) = 0$
- (ii) Si dado $H, A \subset B$, entonces $p(A|H) \leq p(B|H)$
- (iii) $p(\Omega|H) = 1$
- (iv) $p(\bar{A}|H) = 1 - p(A|H)$

Demostración

(i) $p(\Phi|H) = p(\Phi \cup \Phi|H) = p(\Phi|H) + p(\Phi|H) = 2p(\Phi|H)$ en virtud de P2, pero el único número $x, 0 \leq x \leq 1$, que satisface la ecuación $x = 2x$ es $x = 0$.

(ii) Puesto que

$A \cap H \subset B \cap H, B \cap H = (A \cap H) \cup (B \cap \bar{A} \cap H)$ y $A \cap (B \cap \bar{A}) = \Phi$ (véase figura), entonces, en virtud de P2

$$p(B \cap H|H) = p(A \cap H|H) + p(B \cap \bar{A} \cap H|H)$$

y puesto que, en virtud de P1,

$$p(B \cap \bar{A} \cap H|H) \geq 0,$$

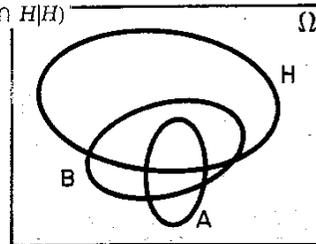
$$p(B \cap H|H) \geq p(A \cap H|H)$$

pero

$$p(B \cap H|H) = p(H|B, H) = p(B|H)$$

y, por tanto,

$$p(B|H) \geq p(A|H)$$



(iii) $\Omega \supset H \cap \Omega = H$ luego, en virtud de (ii) $p(\Omega|H) \geq p(H|H) = 1$ pero como en virtud de P1, $p(\Omega|H) \leq 1$, tenemos $p(\Omega|H) = 1$.

(iv) $p(A \cup \bar{A}|H) = p(\Omega|H) = 1$ en virtud de (iii) y $p(A \cup \bar{A}|H) = p(A|H) + p(\bar{A}|H)$, luego $p(\bar{A}|H) = 1 - p(A|H)$.

La propiedad P2 se refiere únicamente a sucesos mutuamente excluyentes en las condiciones H . Sin embargo, no es difícil generalizarla a sucesos cualesquiera. En efecto,

TEOREMA 3.2.4. Para todo par de sucesos A y B y para cualesquiera condiciones H ,

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

Demostración

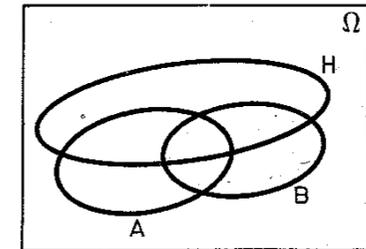
Como se observa en la figura, $A \cup B = (A \cap \bar{B}) \cup (A \cap B) \cup (\bar{A} \cap B)$ y los sucesos $A \cap \bar{B}, A \cap B$ y $\bar{A} \cap B$ son mutuamente excluyentes. En consecuencia, en virtud del Teorema 3.2.1, $p(A \cup B|H) = p(A \cap \bar{B}|H) + p(A \cap B|H) + p(\bar{A} \cap B|H)$.

Pero además, como $A = (A \cap \bar{B}) \cup (A \cap B)$ y $B = (B \cap \bar{A}) \cup (A \cap B)$, son descomposiciones de A y B en conjuntos disjuntos

$$p(B|H) = p(B \cap \bar{A}|H) + p(A \cap B|H)$$

$$p(A|H) = p(A \cap \bar{B}|H) + p(A \cap B|H)$$

El teorema es ahora consecuencia de estas relaciones, sustituyéndolas en la última expresión del párrafo anterior.



El Teorema 3.2.4 puede generalizarse por inducción a un número finito cualquiera de sucesos. Así, para tres sucesos, puede demostrarse que

$$p(A \cup B \cup C|H) = p(A|H) + p(B|H) + p(C|H) - [p(A \cap B|H) + p(B \cap C|H) + p(A \cap C|H)] + p(A \cap B \cap C|H) \quad (2)$$

Ejemplo 3.2.1. Problema del cumpleaños

Determinar la probabilidad p de que al menos dos personas en un grupo de k elegidas al azar tengan el mismo cumpleaños, esto es, que hayan nacido el mismo día del mismo mes, pero no necesariamente del mismo año.

Empezaremos determinando la probabilidad de que los cumpleaños de las k personas sean todos distintos. Cualquiera que sea el cumpleaños de la primera persona, el de la segunda podrá elegirse entre los 364 días restantes, el de la tercera podrá elegirse entre 363 días y así sucesivamente. En consecuencia, la probabilidad de que los k cumpleaños sean distintos será (Teorema 3.2.2),

$$\frac{364}{365} \frac{363}{365} \frac{362}{365} \dots \frac{365-k+1}{365} = \left\{ \prod_{i=1}^{k-1} (365-i) \right\} / 365^{k-1}$$

y, por tanto, la probabilidad pedida, de que al menos dos tengan el mismo cumpleaños será, utilizando el Teorema 3.2.3 (iv),

$$p = 1 - \left\{ \prod_{i=1}^{k-1} (365-i) \right\} / 365^{k-1}$$

Efectuados los cálculos se obtiene para distintos valores de k

k	p	k	p
10	0,117	25	0,569
20	0,411	40	0,891
23	0,507	60	0,994

De manera que, entre 60 personas, es casi seguro que al menos dos tienen el mismo cumpleaños: un resultado difícil de anticipar con razonamientos intuitivos.

En algunos problemas de probabilidades es necesario calcular el número de subconjuntos distintos de k elementos que pueden escogerse de un conjunto de n sin repetirse ninguno y de forma que dos grupos sean distintos si tienen algún elemento diferente. A este número, que se denomina *número combinatorio*, se le representa por $\binom{n}{k}$ y su valor se obtiene de la expresión

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (3)$$

donde $n!$ (leído *factorial* de n) se define como $0! = 1$ y

$$n! = \prod_{i=0}^{n-1} (n-i) = n(n-1)(n-2) \dots \times 2 \times 1, \quad n \geq 1$$

En efecto, los k elementos pueden escogerse en un orden determinado de $n(n-1) \dots (n-k+1)$ maneras distintas: el primero se puede escoger entre n posiciones, el segundo entre $(n-1)$, etc. Por otra parte, en virtud de un razonamiento análogo, existen

$k(k-1) \dots \times 2 \times 1$ formas de ordenar un conjunto de k elementos. En consecuencia, el número de subconjuntos *distintos* de k elementos será

$$\frac{n(n-1) \dots (n-k+1)}{k(k-1) \dots \times 2 \times 1} = \frac{n!}{(n-k)!k!}$$

Ejemplo 3.2.2. Equipos de guardia

Determinar el número de equipos de guardia distintos que pueden formarse en un servicio con 25 médicos 3 de los cuales deben estar de guardia

El número de equipos de guardia distintos será

$$\binom{25}{3} = \frac{25 \times 24 \times 23}{3 \times 2 \times 1} = 2300$$

3.3. Independencia e intercambiabilidad

Sabemos, en virtud del Teorema 2.4.1, que para cualquier par de sucesos A y B ,

$$p(A \cap B|H) = p(A|B, H) p(B|H) = p(B|A, H) p(A|H)$$

Se dice que A y B son *independientes* en las condiciones H si se verifica que

$$p(A \cap B|H) = p(A|H) p(B|H) \quad (1)$$

Esto no es más que un caso particular de la definición siguiente.

DEFINICIÓN 3.3.1. Se dice que los sucesos $\{A_1, A_2, A_3, \dots\}$ son independientes en las condiciones H si para cualquier subconjunto finito de ellos $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ se verifica que

$$p(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}|H) = \prod_{j=1}^k p(A_{i_j}|H)$$

El motivo de la definición de independencia dada por la ecuación (1) es sencillo; en muchas ocasiones la ocurrencia de B puede no afectar a la probabilidad de A en las condiciones H , de forma que

$$p(A|B, H) = p(A|H) \quad (2)$$

lo que podemos enunciar diciendo que A es independiente de \bar{B} . En esta forma no tenemos una definición simétrica, en la que A y B jueguen el mismo papel, pero si multiplicamos ambos miembros de la ecuación (2) por $p(B|H)$ y utilizamos el Teorema 2.4.1, obtenemos la ecuación (1) que es simétrica en A y B .

La extensión a más de dos sucesos exige tener cuidado. Si decimos que los sucesos de un conjunto son independientes entre sí, esto debe interpretarse como que la ocurrencia de un número cualquiera de ellos no afecta a la probabilidad de que ocurra cualquier otro, es decir

$$p(A_i_1|A_i_2, \dots, A_i_k, H) = p(A_i_1|H) \quad (3)$$

ecuación cuya forma simétrica es la Definición 3.3.1, como es fácil de probar.

TEOREMA 3.3.1. *Si A y B son independientes en las condiciones H , los pares A y B , \bar{A} y B , \bar{A} y \bar{B} son también independientes en las mismas condiciones.*

El conjunto A puede partiarse en conjuntos disjuntos en la forma $A = (A \cap B) \cup (A \cap \bar{B})$. En consecuencia, $p(A \cap \bar{B}|H) = p(A|H) - p(A \cap B|H)$, y como A y B son independientes,

$$\begin{aligned} p(A \cap \bar{B}|H) &= p(A|H) - p(A \cap B|H) = \\ &= p(A|H) [1 - p(B|H)] = p(A|H) p(\bar{B}|H) \end{aligned}$$

y por tanto A y \bar{B} son independientes. Utilizando este resultado, y cambiando los papeles de A y B , tenemos que los pares \bar{A} y B y \bar{A} y \bar{B} son asimismo independientes.

Un caso importante de independencia aparece en las llamadas *sucesiones de Bernoulli*. Se dice que una colección de sucesos $\{A_i\}$ forman una sucesión de Bernoulli si la probabilidad de que suceda uno cualquiera de ellos es una constante p independiente de los demás, esto es si

$$p(A_i|H, A_1, A_2, \dots, A_{i-1}, A_{i+1}, \dots) = p(A_i|H) = p \quad (4)$$

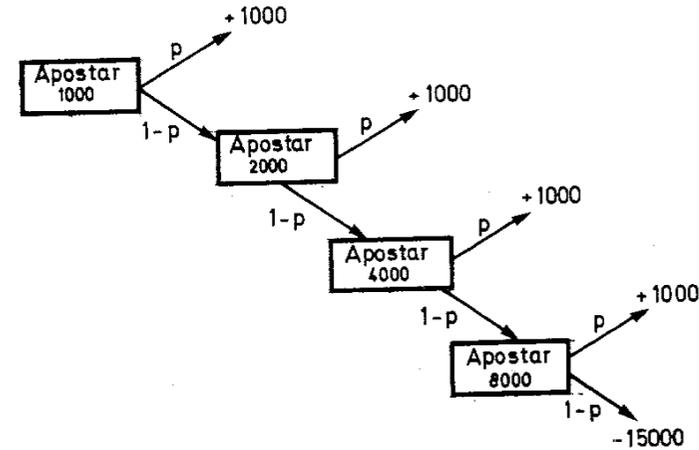
Los juegos de azar proporcionan numerosos ejemplos de sucesiones de Bernoulli.

Ejemplo 3.3.1. Ruleta

Una ruleta está dividida en 37 sectores iguales, 18 de los cuales son rojos, 18 negros y queda uno para la banca. Supongamos que, disponiendo de 15 000 ptas., se juega a la ruleta, a doble o nada, hasta ganar 1.000 ptas. o

quedarse sin dinero, empezando por apostar 1.000. ¿Cuál es la probabilidad de ganar las 1 000 antes de perderlo todo?

Los sucesos relevantes $\{A_i, i = 1, 2, \dots\}$ forman una sucesión de Bernoulli con $A_i = \{\text{Se gana la apuesta } i\}$ y $p(A_i|H) = 18/37 = p$. Obsérvese que, en virtud de la independencia, es indiferente apostar siempre a rojo, siempre a negro, o ir cambiando en cualquier orden. Los posibles desarrollos del juego pueden describirse mediante el árbol siguiente



La probabilidad de perder las 15.000 antes de ganar las 1.000 es pues $(1-p)^4 = 0,0695$, y por tanto la de ganar las 1.000 antes de perderlo todo $1 - (1-p)^4 = 0,9305$.

El ejemplo anterior plantea un caso particular de un problema general muy frecuente en la práctica. Consideremos una sucesión de Bernoulli $\{A_i, i = 1, 2, \dots\}$ y supongamos que $p(A_i|H) = p$, y por lo tanto $p(\bar{A}_i|H) = 1 - p$. Consideremos n sucesos cualquiera de la sucesión y tratemos de determinar la probabilidad de que ocurran precisamente k de ellos, $0 \leq k \leq n$. Como todos los sucesos A_i son independientes, la probabilidad de que ocurran k de ellos, y dejen de ocurrir por lo tanto $n - k$, será el producto de sus probabilidades respectivas, esto es $p^k(1-p)^{n-k}$. Como existen $\binom{n}{k}$ sucesiones ordenadas de este tipo, mutuamente excluyentes (ver Ejemplo 3.2.2) la probabilidad buscada es

$$p(k|n, p, H) = \binom{n}{k} p^k(1-p)^{n-k} \quad (5)$$

Para cada n , los valores de k pueden ir de 0 a n ambos inclusive y,

puesto que se trata de sucesos mutuamente excluyentes, la suma de las $n + 1$ probabilidades obtenidas deberá ser necesariamente la unidad, como puede comprobarse utilizando la fórmula del binomio de Newton.

Ejemplo 3.3.2. Sexo de un recién nacido

Consideremos los n próximos alumbramientos que tengan lugar en el Hospital Clínico. Determinar la probabilidad de que k de ellos den lugar a niñas.

Los sucesos relevantes $\{A_i, i = 1, 2, \dots\}$ forman una sucesión de Bernoulli con $A_i = \{\text{niña}\}$ y $\bar{A}_i = \{\text{niño}\}$. Los datos de que se dispone permiten asegurar que, en condiciones normales H , esto es, en ausencia de información que permita asegurar que alguna de las madres tenga una tendencia distinta de la normal a tener hijos varones, $p(A_i|H) = 0,49$. En consecuencia, la probabilidad pedida es $p(k \text{ niñas} | n \text{ partos}, H) =$

$$= \binom{n}{k} 0,49^k \times 0,51^{n-k}$$

Si, por ejemplo, $n = 3$ se obtiene

$$p(k = 0|H) = 0,1327$$

$$p(k = 1|H) = 0,3823$$

$$p(k = 2|H) = 0,3675$$

$$p(k = 3|H) = 0,1175$$

Estas son también las probabilidades de las distintas composiciones de la familia de una pareja que decida tener tres hijos, en ausencia de información que permita asegurar una tendencia distinta de la normal a tener hijos de uno u otro sexo.

En una sucesión de Bernoulli, $\{A_i, i = 1, 2, \dots\}$ los sucesos que la constituyen son, por definición, independientes *dado* p , de forma que, si la probabilidad $p(A_i|H) = p$ de que suceda uno cualquiera de ellos es conocida, la probabilidad de que ocurran r de un conjunto de n en un orden determinado es $p^r(1-p)^{n-r}$ independientemente del orden elegido. Sin embargo, si el valor de p es desconocido, los sucesos A_i ya no son independientes; en efecto, la ocurrencia o no ocurrencia de uno de ellos proporciona información sobre el valor común y desconocido de p y por lo tanto influye en la probabilidad asignada a los demás. Puede demostrarse no obstante que la probabilidad de que ocurran r de un conjunto de n en un orden determinado sigue siendo independiente del orden elegido. Diremos entonces que los sucesos $\{A_i, i = 1, 2, \dots\}$ son *intercambiables*. Formalmente

DEFINICIÓN 3.3.2. Una colección $\{A_i, i = 1, 2, \dots\}$ de sucesos son intercambiables si para todo subconjunto finito $\{A_1, A_2, \dots, A_n\}$, la probabilidad

de que ocurran r de los n en un orden determinado es independiente del orden elegido y del subconjunto considerado.

Ovviamente, una sucesión de Bernoulli $\{A_i, i = 1, 2, \dots\}$ con $p(A_i|H) = p$ conocido es una colección de sucesos intercambiables, puesto que la probabilidad de que ocurran r de un conjunto de n es $p^r(1-p)^{n-r}$, independientemente del orden elegido y del conjunto considerado. En el ejemplo próximo comprobaremos que una sucesión de Bernoulli con $p(A_i|H) = p$ desconocido, sigue siendo una colección de sucesos intercambiables.

Ejemplo 3.3.3. Aparición de tumores

Una determinada enfermedad puede presentarse en dos formas, benigna y maligna. En la forma maligna, da lugar a tumor cerebral con probabilidad 0,7, mientras que en la forma benigna esta probabilidad es tan solo 0,1. Se sabe además que la forma benigna es 4 veces más probable que la maligna. Sea A_i el suceso consistente en que aparezca un tumor en la persona i que padece la enfermedad. Demostrar que los sucesos $\{A_i, i = 1, 2, \dots\}$ son intercambiables pero no independientes.

Sea B la forma benigna y $M = \bar{B}$ la forma maligna de la enfermedad. Omitiendo, por simplificar la notación, las condiciones generales H , sabemos que

$$p(A_i|B) = 0,1, \quad p(A_i|M) = 0,7,$$

$$p(B) = 4p(M) = 4(1 - p(B)) \rightarrow p(B) = 0,8$$

Puesto que $A_i \cap B$ y $A_i \cap M$ son disjuntos y su unión es A_i ,

$$p(A_i) = p(A_i \cap B) + p(A_i \cap M) = p(A_i|B) p(B) + p(A_i|M) p(M)$$

en virtud de las leyes aditiva y multiplicativa (teorema 2.4.1) de la probabilidad. En consecuencia,

$$p(A_i) = 0,1 \times 0,8 + 0,7 \times 0,2 = 0,22$$

Análogamente,

$$\begin{aligned} p(A_i \cap A_j) &= p(A_i \cap A_j|B) p(B) + p(A_i \cap A_j|M) p(M) \\ &= (0,1)^2 0,8 + (0,7)^2 0,2 = 0,106 \end{aligned}$$

Claramente,

$$p(A_i \cap A_j) = 0,106 \neq p(A_i) p(A_j) = 0,22^2 = 0,0484$$

y, en consecuencia, los sucesos $\{A_i, i = 1, 2, \dots\}$ no son independientes. Sin embargo la probabilidad de que entre n pacientes aparezcan k tumores es

$$p(k|n) = p(k|n, B) p(B) + p(k|n, M) p(M) \\ = \left\{ \binom{n}{k} 0,1^k \times 0,9^{n-k} \right\} 0,8 + \left\{ \binom{n}{k} 0,7^k \times 0,3^{n-k} \right\} 0,2$$

que es independiente del orden en que se presenten, de forma que los $\{A_i, i = 1, 2, \dots\}$ son intercambiables.

3.4. Teoremas de Bayes y de la probabilidad total

Como hemos ilustrado en el Ejemplo 3.3.3, la forma más sencilla de determinar la probabilidad de un suceso se basa frecuentemente en la consideración de la probabilidad de ese suceso bajo un conjunto de condiciones mutuamente excluyentes. Así, en el ejemplo citado, las probabilidades de que apareciesen tumores fueron obtenidas considerando las probabilidades de que apareciesen en las condiciones, mutuamente excluyentes, de que la enfermedad fuese benigna o que fuese maligna. El resultado que vamos a demostrar a continuación, conocido como teorema de la *probabilidad total* y también como *tercera ley* de la probabilidad, pone de manifiesto la relación entre la probabilidad de un suceso y las probabilidades de ese mismo suceso en condiciones mutuamente excluyentes.

TEOREMA 3.4.1. (Teorema de la probabilidad total.) *Para todo suceso $A \subset \Omega$, condiciones H y conjunto de sucesos mutuamente excluyentes B_1, B_2, \dots, B_k tales que $\Omega = \cup B_i$*

$$p(A|H) = \sum_{i=1}^k p(A|B_i, H) p(B_i|H)$$

Demostración

Si $B_i \cap B_j = \Phi$ y $\Omega = \cup B_i$, entonces los sucesos del tipo $A \cap B_i$ son disjuntos y $A = \cup \{A \cap B_i\}$. Por lo tanto, en virtud de la ley aditiva

$$p(A|H) = \sum_{i=1}^k p(A \cap B_i|H)$$

y puesto que, en virtud de la ley multiplicativa, $p(A \cap B_i|H) = p(A|B_i, H) p(B_i|H)$ tenemos el resultado buscado.

El teorema de la probabilidad total permite a menudo expresar la probabilidad del suceso incierto en que estamos interesados en función de otras probabilidades más sencillas de obtener.

Ejemplo 3.4.1. Premio literario

En un determinado premio literario, se anuncian los nombres de los tres finalistas A, B y C , pero se reserva el nombre del ganador para más tarde. El concursante A razona entonces al presentador que puesto que se sabe que, necesariamente, al menos uno de los otros dos B o C , tiene que ser eliminado no le da ninguna información sobre sus propias posibilidades si le dice el nombre de uno de los otros dos que haya sido eliminado. ¿Es correcta su argumentación?

Sean A, B, C respectivamente los sucesos consistentes en que A, B o C sea el ganador: supongamos que al presentador le convence el argumento del concursante A y le confía, sin mentir, que B ha sido eliminado. Si denotamos b este suceso, nuestro problema se reduce a determinar $p(A|b)$ y comprobar si esta probabilidad es o no es igual a $p(A)$. Utilizando el Teorema 3.4.1,

$$p(b) = p(b|B) p(A) + p(b|B) p(B) + p(b|C) p(C) \\ = p(b|A) p(A) + 0 + p(C)$$

y en virtud de la ley multiplicativa

$$p(A \cap b) = p(A|b) p(b) \\ p(A \cap b) = p(b|A) p(A)$$

de forma que

$$p(A|b) = p(b|A) p(A) / p(b) = \frac{p(b|A)}{p(b|A) p(A) + p(C)} p(A)$$

por lo tanto, $p(A|b)$ será igual a $p(A)$, como el concursante A razonaba al presentador si, y solamente si, $p(b|A) = p(b|A) p(A) + p(C)$, esto es si $p(b|A) = p(C) / (1 - p(A))$. Si, como parece razonable suponer, $p(b|A) = 0,5$ esto es si se juzga que en el caso de ser A el ganador el presentador nombraría a B o a C con la misma probabilidad, entonces la información b dada por el presentador sería irrelevante para A si, $p(A) = 1 - 2p(C)$. Esto sucede si, y solamente si, $p(B) = p(C)$, esto es siempre que se juzgue que B y C tienen inicialmente la misma probabilidad de ganar. En consecuencia, si $p(b|A) = 0,5$, la información dada por el presentador es verdaderamente irrelevante para A si en su opinión sus dos contrincantes tienen la misma probabilidad de alcanzar el premio y no lo es en caso contrario.

La reacción natural en cualquiera que tenga que tomar una decisión en ambiente de incertidumbre es tratar de reducirla incorporando cuanta información adicional sobre los sucesos inciertos le sea posible.

La información del decisor sobre un conjunto de sucesos inciertos exhaustivos y mutuamente excluyentes $\theta_1, \theta_2, \dots, \theta_k$ se describe, según vimos en el Capítulo 2, mediante unas probabilidades $p(\theta_1|H), p(\theta_2|H), \dots, p(\theta_k|H)$ tales que $0 \leq p(\theta_i|H) \leq 1$ y $\sum p(\theta_i|H) = 1$. El efecto de la información adi-

cional será el de modificar o revisar estas probabilidades. Si se llega a obtener una información completa, una de estas probabilidades se hará igual a 1 y el resto se harán igual a 0. Una información incompleta producirá cambios menos acentuados. Si llamamos X a esta información adicional, los nuevos valores de las probabilidades asignadas a los sucesos inciertos, después de haber obtenido la información X , serán $p(\theta_1|X, H)$, $p(\theta_2|X, H)$, ..., $p(\theta_k|X, H)$. El resultado que vamos a demostrar a continuación, conocido como *Teorema de Bayes* sirve para obtener las probabilidades *finales* $p(\theta_i|X, H)$ a partir de las probabilidades *iniciales* $p(\theta_i|H)$ y de la relación que exista entre la información obtenida X y los sucesos inciertos θ_i .

El Teorema de Bayes es en realidad una consecuencia inmediata de la ley multiplicativa de la probabilidad, cuyo uso para determinar la probabilidad de un suceso, dada cierta información adicional, hemos ilustrado ya en el Ejemplo 3.4.1.

TEOREMA 3.4.2 (Teorema de Bayes.) *Para todo par de sucesos A, B y para cualesquiera condiciones H ,*

$$p(A|B, H) = p(B|A, H) p(A|H)/p(B|H)$$

Demostración

En virtud de la ley multiplicativa de la probabilidad,

$$p(A \cap B|H) = p(A|H) p(B|A, H)$$

y también

$$p(A \cap B|A) = p(B|H) p(A|B, H)$$

Iguando ambas ecuaciones y despejando $p(A|B, H)$ se obtiene el resultado enunciado.

Consecuentemente, en un problema de decisión, las probabilidades *finales* de los sucesos inciertos, $\theta_1, \theta_2, \dots, \theta_k$ después de haber obtenido cierta información adicional X vendrán dadas por

$$p(\theta_i|X, H) = p(X|\theta_i, H) p(\theta_i|H)/p(X|H), \quad i = 1, 2, \dots, k \quad (1)$$

Puesto que la probabilidad $p(X|H)$ que aparece en el segundo miembro de (1) es común a todos los sucesos inciertos podemos escribir el resultado en forma *proporcional* como

$$p(\theta_i|X, H) \propto p(X|\theta_i, H) p(\theta_i|H) \quad (2)$$

donde el símbolo \propto se lee «proporcional a». En efecto, a partir de la ex-

presión (2) siempre pueden obtenerse las probabilidades deseadas puesto que los valores de las $p(\theta_i|X, H)$ deben sumar uno. Naturalmente, el valor de la constante de proporcionalidad $1/p(X|H)$ puede deducirse directamente puesto que, en virtud del Teorema de la probabilidad total,

$$p(X|H) = \sum_{i=1}^k p(X|\theta_i, H) p(\theta_i|H) \quad (3)$$

Ejemplo 3.4.2. Test de tuberculina

En una campaña de erradicación de la tuberculosis, se somete al test de la tuberculina a los alumnos de todos los centros de Enseñanza Media. Se sabe que la probabilidad de que una persona que tiene tuberculosis dé lugar a un test positivo es 0,98, mientras que la probabilidad de que el test dé positivo en una persona sana es 0,05. Sabiendo que un 1 % de los alumnos a quienes se aplica el test padecen tuberculosis, determinar la probabilidad de que un alumno escogido al azar, padezca tuberculosis en función del resultado del test de tuberculina.

Sean + y - los posibles resultados del test y sea T el suceso de que el alumno tenga tuberculosis. Omitiendo por comodidad de notación las condiciones H , sabemos que $p(+|T) = 0,98$ y $p(+|\bar{T}) = 0,05$, esto es los porcentajes de error del test son del 2 % en caso positivo y del 5 % en caso negativo, y sabemos también que $p(T) = 0,01$, esto es que la probabilidad *inicial* de que un alumno tenga tuberculosis es 0,01. Deseamos determinar $p(T|+)$ y $p(T|-)$. Utilizando (2),

$$p(T|+) \propto p(+|T) p(T) = 0,98 \times 0,01 = 0,0098$$

$$p(\bar{T}|+) \propto p(+|\bar{T}) p(\bar{T}) = 0,05 \times 0,99 = 0,0495$$

$$0,0593$$

y puesto que $p(T|+) + p(\bar{T}|+) = 1$,

$$p(T|+) = 0,0098/0,0593 = 0,165$$

$$p(\bar{T}|+) = 0,0495/0,0593 = 0,835$$

de forma que si el test da positivo la probabilidad de tuberculosis pasa de 0,01 a 0,165, multiplicándose por 16,5. Obsérvese sin embargo que, a pesar de ser relativamente bajos los porcentajes de error del test, es todavía *cinco* veces más probable (0,835/0,165 = 5,06) que esté sano a que esté tuberculoso un alumno a quien ha dado positivo el test. La razón de ello estriba en la pequeña incidencia (1 %) de la enfermedad.

Análogamente, si el test da negativo

$$p(T|-) \propto p(-|T) p(T) = 0,02 \times 0,01 = 0,0002$$

$$p(\bar{T}|-) \propto p(-|\bar{T}) p(\bar{T}) = 0,95 \times 0,99 = 0,9405$$

$$0,9407$$

de forma que si el test da negativo la probabilidad de que el alumno tenga tuberculosis pasa de 0.01 a $0,0002/0,9407 \approx 0,000213$, dividiéndose por 47.

Una consecuencia inmediata del Teorema de Bayes, es la de que no resulta razonable asignar probabilidades nulas a sucesos inciertos que se consideren poco probables. En efecto, si la probabilidad inicial de un suceso A es $p(A|H) = 0$ entonces, en virtud del Teorema de Bayes, su probabilidad final $p(A|B, H)$ también será cero *cualquiera* que sea la información proporcionada por B . El uso de probabilidades cero o uno representa una actitud dogmática, pocas veces compatible con la investigación científica. Así por ejemplo, en un problema de diagnosis, se podrá asignar una probabilidad muy pequeña a enfermedades que se consideren muy poco probables pero, en general, no sería razonable asignarles probabilidad cero, para cubrir la posibilidad de que los datos clínicos (o eventualmente la necropsia) demostraran que la impresión inicial era incorrecta.

Ejemplo 3.4.3. *Diagnosis*

Un médico duda entre tres enfermedades posibles en un paciente, que denotaremos por E_1, E_2 y E_3 . A la vista del estado general del enfermo, E_1 le parece la causa más probable de la dolencia, tres veces más probable que cualquiera de las otras dos. Sin embargo, ordena un análisis de sangre y resulta que los resultados X del análisis se asocian rara vez con E_1 , muy frecuentemente con E_2 y a menudo con E_3 , de forma que $p(X|E_1) = 0,1$, $p(X|E_2) = 0,9$ y $p(X|E_3) = 0,6$. ¿Cuál es la probabilidad final de cada una de las enfermedades?

Claramente, omitiendo por brevedad las condiciones H ,

$$\left\{ \begin{array}{l} p(E_1) = 3p(E_2) \\ p(E_1) = 3p(E_3) \\ p(E_1) + p(E_2) + p(E_3) = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} p(E_1) = 0,6 \\ p(E_2) = 0,2 \\ p(E_3) = 0,2 \end{array} \right.$$

Por otra parte, en virtud del Teorema de Bayes,

$$\begin{aligned} p(E_1|X) &\propto p(X|E_1) \quad p(E_1) = 0,1 \times 0,6 = 0,06 \\ p(E_2|X) &\propto p(X|E_2) \quad p(E_2) = 0,9 \times 0,2 = 0,18 \\ p(E_3|X) &\propto p(X|E_3) \quad p(E_3) = 0,6 \times 0,2 = \frac{0,12}{0,36} \end{aligned}$$

y, en consecuencia,

$$\begin{aligned} p(E_1|X) &= 0,06/0,36 = 1/6 \\ p(E_2|X) &= 0,18/0,36 = 1/2 \\ p(E_3|X) &= 0,12/0,36 = 1/3 \end{aligned}$$

y, por tanto, *tras* observar los resultados del análisis la enfermedad E_1 resulta tres veces *menos* probable que E_3 como causa de la afección.

3.5. Análisis secuencial

Cuando la información adicional sobre los sucesos inciertos se obtiene de forma secuencial, el Teorema de Bayes puede ser aplicado sucesivamente utilizando como probabilidades iniciales en cada fase las probabilidades finales de la fase anterior.

Supongamos por ejemplo que se desea obtener información sobre un determinado suceso incierto A , para lo que se realizan dos experimentos cuyos resultados son B_1 y B_2 . Entonces, según el Teorema de Bayes,

$$p(A|B_1, B_2, H) = p(B_1, B_2|A, H) p(A|H) / p(B_1, B_2|H) \quad (1)$$

pero, en virtud de la ley multiplicativa de la probabilidad

$$p(B_1, B_2|A, H) = p(B_1|A, H) p(B_2|B_1, A, H) \quad (2)$$

$$p(B_1, B_2|H) = p(B_1|H) p(B_2|B_1, H) \quad (3)$$

y sustituyendo en (1)

$$\begin{aligned} p(A|B_1, B_2, H) &= \frac{p(B_2|B_1, A, H)}{p(B_2|B_1, H)} \frac{p(B_1|A, H)}{p(B_1|H)} p(A|H), \\ &= \frac{p(B_2|A, B_1, H)}{p(B_2|B_1, H)} p(A|B_1, H) \end{aligned} \quad (4)$$

puesto que según el Teorema de Bayes,

$$p(A|B_1, H) = p(B_1|A, H) p(A|H) / p(B_1|H) \quad (5)$$

pero la expresión (4) es de nuevo el Teorema de Bayes, incorporando la información B_2 una vez conocida la información B_1 : el conjunto (B_1, H) hace en (4) el papel que H hace en (5) y B_2 juega en (4) el papel que B_1 juega en (5).

Ejemplo 3.5.1. *Prueba del alcohol*

En muchos países existe una ley que relaciona el derecho a conducir con el contenido de alcohol en la sangre. Un test rápido, realizado por la policía

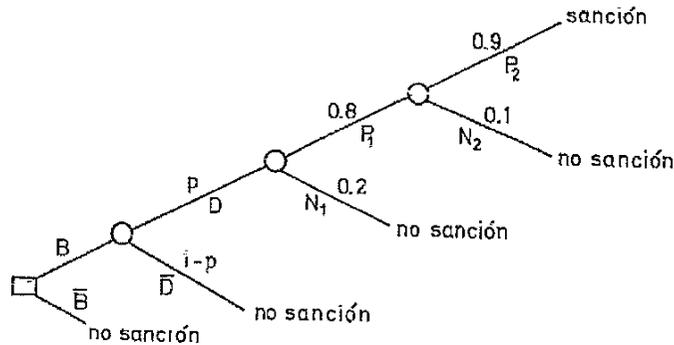
con un aparato portátil, tiene solamente una probabilidad 0,8 de ser correcto entre las personas que la policía controla, esto es, de dar positivo si el contenido de alcohol en la sangre está por encima del límite legal y de dar negativo en caso contrario. Aquellos conductores a quienes el test efectuado por la policía les da positivo son sometidos en la comisaría a un test más riguroso efectuado por un médico. Este segundo test nunca da resultados incorrectos con un conductor sobrio, pero da negativo en el 10 % de los conductores detenidos que habían, de hecho, sobrepasado el límite legal, debido al tiempo transcurrido desde su detención. Suponiendo ambos tests independientes, determinar la probabilidad de que un conductor que ha sobrepasado el límite legal no sea sancionado, en función de la probabilidad p de que sea controlado por la policía cuando ha bebido. Suponiendo que la probabilidad de ser controlado por la policía es 0,5 cuando se ha bebido más de lo legal y 0,1 en caso contrario, y que el 15 % de los conductores conducen con más alcohol en la sangre del permitido legalmente, determinar la probabilidad de que un conductor haya bebido más de lo legalmente permitido en función de los resultados de los tests.

Omitiendo por sencillez de notación las condiciones generales H del problema, denotando por B el suceso de que el conductor haya bebido más de lo permitido, por P_1, N_1, P_2, N_2 los de que el primer y el segundo test respectivamente den positivos o negativos, y por D el suceso de que el conductor sea detenido por la policía y sometido al primer test,

$$p(P_1|D, B) = 0,8, \quad p(P_1|D, \bar{B}) = 0,2$$

$$p(P_2|D, B) = 0,9, \quad p(P_2|D, \bar{B}) = 0$$

y $p(D|B) = p$. La situación puede describirse mediante un árbol en la forma siguiente



La probabilidad de que un conductor que ha sobrepasado el límite legal sea sancionado es pues $p \times 0,8 \times 0,9 = 0,72p$ y por tanto la de no serlo $1 - 0,72p$. Si, por ejemplo,

existe una probabilidad del 50 % de ser controlado por la policía cuando se ha bebido, entonces existe una probabilidad 0,64 de no ser sancionado.

Si sabemos que $p(D|B) = 0,5$, $p(D|\bar{B}) = 0,1$ y $p(B) = 0,15$, entonces, por Teorema de Bayes,

$$p(B|D) \propto p(D|B) p(B) = 0,5 \times 0,15 = 0,075$$

$$p(\bar{B}|D) \propto p(D|\bar{B}) p(\bar{B}) = 0,1 \times 0,85 = 0,085$$

$$\hline 0,160$$

de forma que $p(B|D) = 0,075/0,16 = 0,469$ y $p(\bar{B}|D) = 0,531$. De forma análoga, se encuentra $p(B|\bar{D}) = 0,089$.

Aplicando de nuevo el Teorema de Bayes,

$$p(B|D, P_1) \propto p(P_1|D, B) p(B|D) = 0,8 \times 0,469 = 0,3752$$

$$p(\bar{B}|D, P_1) \propto p(P_1|D, \bar{B}) p(\bar{B}|D) = 0,2 \times 0,531 = 0,1062$$

$$\hline 0,4814$$

de forma que $p(B|D, P_1) = 0,3752/0,4814 = 0,779$ y $p(\bar{B}|D, P_1) = 0,221$. De manera totalmente análoga, se encuentra que

$$p(B|\bar{D}, N_1) = 0,181 \quad p(\bar{B}|\bar{D}, N_1) = 0,819$$

Si el primer test da negativo el segundo va o no se realiza; si el primer test da positivo tenemos, aplicando una vez más el Teorema de Bayes,

$$p(B|D, P_1, P_2) \propto p(P_2|P_1, D, B) p(B|D, P_1)$$

$$p(\bar{B}|D, P_1, P_2) \propto p(P_2|P_1, D, \bar{B}) p(\bar{B}|D, P_1)$$

pero como los tests son independientes, $p(P_2|P_1, D, B) = p(P_2|D, B)$ y $p(P_2|P_1, D, \bar{B}) = p(P_2|D, \bar{B})$, de forma que

$$p(B|D, P_1, P_2) \propto 0,9 \times 0,779 = 0,7011$$

$$p(\bar{B}|D, P_1, P_2) \propto 0 \times 0,221 = 0$$

$$\hline 0,7011$$

y, por tanto $p(B|D, P_1, P_2) = 1$ y $p(\bar{B}|D, P_1, P_2) = 0$.

Análogamente, si el segundo test da negativo,

$$p(B|\bar{D}, P_1, N_2) \propto p(N_2|\bar{D}, B) p(B|\bar{D}, P_1) = 0,1 \times 0,779 = 0,0779$$

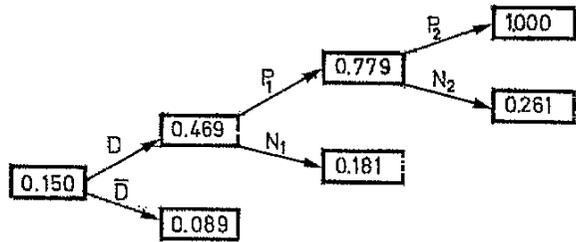
$$p(\bar{B}|\bar{D}, P_1, N_2) \propto p(N_2|\bar{D}, \bar{B}) p(\bar{B}|\bar{D}, P_1) = 1 \times 0,221 = 0,2210$$

$$\hline 0,2989$$

puesto que $p(N_2|D, B) = 1 - p(P_2|D, B)$ y análogamente $p(N_2|\bar{D}, \bar{B}) = 1 - p(P_2|\bar{D}, \bar{B})$. En consecuencia,

$$p(B|D, P_1, N_2) = 0,0779/0,2989 = 0,261 \quad \vee \quad p(\bar{B}|D, P_1, N_2) = 0,739$$

Los resultados obtenidos pueden resumirse en el siguiente diagrama que describe los cambios que sufre la probabilidad de B, esto es la probabilidad de que el conductor tenga un contenido de alcohol en la sangre mayor del legalmente permitido, en función de la información proporcionada por el hecho de ser detenido y por los sucesivos tests a que es entonces sometido



Ejemplo 3.5.2. Valor de la información

Una empresa farmacéutica se plantea la posibilidad de lanzar al mercado un nuevo antigripal. Los beneficios que puede obtener dependen de la proporción θ de médicos que lo recetan. Un equipo de marketing les ofrece la realización de una encuesta para reducir la incertidumbre sobre θ . Simplificando el problema, puede suponerse que la proporción θ de médicos que recetarían el nuevo antigripal puede ser alta (θ_1), media (θ_2) o baja (θ_3) y que los beneficios que la empresa puede esperar son, respectivamente, 5, 1 y -3 millones de pesetas. La información inicial de la empresa farmacéutica permite suponer que $p(\theta_1) = 0,2$, $p(\theta_2) = 0,5$, $p(\theta_3) = 0,3$. La encuesta propuesta puede aconsejar la producción ($x = 1$) o desaconsejarla ($x = 0$) y las probabilidades de que el resultado de la encuesta sea aconsejar la producción, en función de θ , son $p(x=1|\theta_1) = 0,9$, $p(x=1|\theta_2) = 0,5$ y $p(x=1|\theta_3) = 0,2$. Suponiendo la utilidad proporcional al dinero, determinar el precio máximo que debe pagarse por la encuesta.

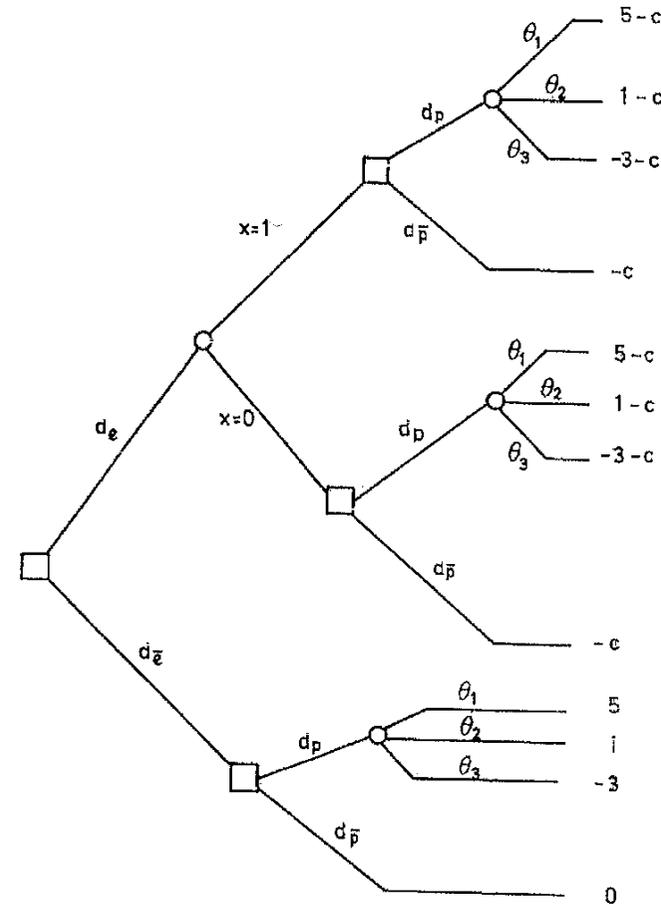
Si denotamos por d_e la decisión de encargar la encuesta, por a_p la de producir el nuevo antigripal y por c el coste de la encuesta, la situación puede ser descrita mediante el árbol de la figura. Así, si se acepta la encuesta, y esta produce un resultado positivo ($x = 1$), pueden ganarse $5 - c$, $1 - c$ o $-3 - c$ millones, con probabilidades $p(\theta_1|x = 1)$, $p(\theta_2|x = 1)$ y $p(\theta_3|x = 1)$ respectivamente. Estas probabilidades pueden calcularse mediante el Teorema de Bayes. En efecto,

$$p(\theta_1|x = 1) \propto p(x = 1|\theta_1) p(\theta_1) = 0,9 \times 0,2 = 0,18$$

$$p(\theta_2|x = 1) \propto p(x = 1|\theta_2) p(\theta_2) = 0,5 \times 0,5 = 0,25$$

$$p(\theta_3|x = 1) \propto p(x = 1|\theta_3) p(\theta_3) = 0,2 \times 0,3 = 0,06$$

$$0,49 = p(x = 1)$$



y, por tanto,

$$p(\theta_1|x = 1) = 0,367 \quad p(\theta_2|x = 1) = 0,510 \quad p(\theta_3|x = 1) = 0,123$$

de forma que la utilidad esperada de producir (d_p) cuando el resultado de la encuesta lo aconseja ($x = 1$) es

$$u^*(d_p|x=1) = 0,367(5-c) + 0,510(1-c) + 0,123(-3-c) = 1,976 - c$$

Obviamente, la utilidad esperada de no producir (d_n) cuando el resultado de la encuesta es $x = 1$ es

$$u^*(d_n|x=1) = -c$$

En consecuencia, si el resultado de la encuesta es $x = 1$ debe tomarse la decisión de producir, y se tiene una utilidad esperada de $1,976 - c$ millones.

Análogamente si el resultado de la encuesta es $x = 0$, tenemos que

$$p(\theta_1|x=0) \text{ cc } p(x=0|\theta_1) \cdot p(\theta_1) = 0,1 \times 0,2 = 0,02$$

$$p(\theta_2|x=0) \text{ cc } p(x=0|\theta_2) \cdot p(\theta_2) = 0,5 \times 0,5 = 0,25$$

$$p(\theta_3|x=0) \text{ cc } p(x=0|\theta_3) \cdot p(\theta_3) = 0,8 \times 0,3 = 0,24$$

$$0,51 = p(x=0)$$

y por tanto

$$p(\theta_1|x=0) = 0,039 \quad p(\theta_2|x=0) = 0,490 \quad p(\theta_3|x=0) = 0,471$$

de forma que la utilidad esperada de producir (d_p) cuando el resultado de la encuesta lo desaconseja ($x = 0$) es

$$u^*(d_p|x=0) = 0,039(5-c) + 0,490(1-c) + 0,471(-3-c) = -0,728 - c$$

que es menor que $u^*(d_n|x=0) = -c$. Por lo tanto, si el resultado de la encuesta es $x = 0$, la decisión óptima es no producir, con lo que se habrá perdido la cantidad c , que costó la encuesta.

Por otra parte, la probabilidad de que la encuesta aconseje producir, es, en virtud del teorema de la probabilidad total,

$$p(x=1) = p(x=1|\theta_1) \cdot p(\theta_1) + p(x=1|\theta_2) \cdot p(\theta_2) + p(x=1|\theta_3) \cdot p(\theta_3)$$

Esta suma ya ha sido calculada, obteniéndose $p(x=1) = 0,49$. Análogamente $p(x=0) = 0,51$.

En consecuencia, la utilidad esperada si se encarga la encuesta es

$$\begin{aligned} u^*(d_e) &= u^*(d_p|x=1) \cdot p(x=1) + u^*(d_n|x=0) \cdot p(x=0) \\ &= (1,976 - c) \times 0,49 - c \times 0,51 = 0,968 - c \end{aligned}$$

Finalmente, si no se encarga la encuesta, la utilidad esperada de producir es, utilizando las probabilidades iniciales,

$$u^*(d_p) = 0,2 \times 5 + 0,5 \times 1 + 0,3 \times (-3) = 0,6$$

mientras la de no producir es obviamente cero. En consecuencia, la decisión óptima si no se encarga la encuesta es producir, con utilidad esperada de 0,6 millones.

La encuesta debe encargarse si, y solamente si, $u^*(d_e) > u^*(d_p)$ esto es si $0,968 - c > 0,6$ y por tanto $c < 0,368$. Así pues el precio máximo que debe pagarse por la encuesta es 368.000 pesetas. Si el precio c de la encuesta es inferior a 368.000 ptas., la estrategia (sucesión de decisiones) óptima es encargarla, y actuar según su resultado, esto es produciendo si la encuesta lo aconseja y no produciendo en caso contrario, con lo que se tiene una utilidad esperada de $0,968 - c$ millones de pesetas. Si el precio c de la encuesta es superior a 368.000 pesetas, la estrategia óptima es no encargarla y pasar a producir el antigripal, con una utilidad esperada de 0,6 millones de pesetas.

3.6. Asignación de probabilidades

En el capítulo segundo ha sido establecida la necesidad axiomática de cuantificar mediante probabilidades nuestra información inicial sobre la relativa verosimilitud de los sucesos inciertos que pueden afectar a las consecuencias de las decisiones. Se advirtió entonces que, aunque se daba una definición constructiva de la probabilidad, éste no era necesariamente el mejor método de asignar probabilidades. En esta sección nos ocuparemos de desarrollar algunos métodos alternativos.

Las probabilidades reflejan la información del decisor. Si esta información se basa esencialmente en resultados muestrales previos, entonces es razonable suponer que las probabilidades asignadas serán próximas a las frecuencias relativas. Si se basa en la percepción de determinadas simetrías, podemos esperar una relación con la definición clásica de probabilidad. Sin embargo, todas las probabilidades son subjetivas y, en general, la información anterior (que algunos llamaban «objetiva») se combinará con otro tipo de información, muy relevante, pero difícilmente expresable en términos de simetrías o de frecuencias relativas. Así, la probabilidad que un cirujano asigna al éxito de una operación dependerá en parte de la frecuencia relativa con que la operación en cuestión ha tenido éxito en situaciones «similares»; pero es obvio que la declaración de «similaridad» entre determinados pacientes es un juicio subjetivo del cirujano y que, además, la probabilidad que asigne dependerá asimismo, y muy notablemente, de la impresión que tiene el cirujano sobre el caso particular que presenta el paciente que va a operar. El cirujano deberá coordinar ambos tipos de información para asignar de forma razonable la probabilidad que necesita.

Existen muchas técnicas diseñadas para especificar probabilidades. En general, se basan en pedir al decisor que elija entre una serie de opciones alternativas diseñadas de forma que las probabilidades buscadas puedan deducirse de sus respuestas.

Supongamos, por ejemplo, que deseamos obtener la probabilidad $p(A|H)$, que, en las condiciones H , el decisor asigna al suceso A . Una forma de hacerlo consiste en pedir al decisor que, en las condiciones H , elija entre

- (i) quedarse como estaba (*status quo*)
 (ii) recibir X si sucede A o recibir Y en caso contrario.

Modificando convenientemente X e Y obtendremos un par de valores X_0 y Y_0 tales que (i) y (ii) resultan igualmente deseables para el decisor. En tal caso, si representamos por S la situación de *status quo* tendremos, en virtud de los principios de coherencia

$$S \sim \{X|A, Y|\bar{A}\} \quad (1)$$

$$u(S) = u(X) p(A|H) + u(Y) \{1 - p(A|H)\}$$

$$p(A|H) = \frac{u(S) - u(Y)}{u(X) - u(Y)} \quad (2)$$

Un caso particular muy frecuente es que X e Y sean respectivamente un beneficio y una pérdida monetarios. Si las cantidades involucradas son suficientemente pequeñas para que la utilidad pueda considerarse proporcional al dinero, tendremos $u(S) = 0$, $u(X) = aX$ y $u(Y) = -aY$, donde a es una constante de proporcionalidad. Consecuentemente, (2) se reduce a

$$p(A|H) = \frac{0 + aY}{aX + aY} = \frac{Y}{X + Y} \quad (3)$$

que resulta, de hecho, independiente de la constante de proporcionalidad a elegida.

Si el decisor fuese perfectamente coherente, la probabilidad obtenida con otros pares (X_i, Y_i) que satisfagan la condición (1) debería ser la misma. En general, no lo será, pero la media de los valores obtenidos será una estimación razonable de la probabilidad que el decisor asigna al suceso A en las condiciones H .

Ejemplo 3.6.1. Pronóstico de un partido

Una persona encuentra justa una apuesta según la cual recibe 600 ptas. si un determinado partido lo gana el equipo local y debe pagar 1.000 el caso contrario, y una segunda apuesta según la cual recibe 1.200 en caso de empate y debe pagar 400 si el empate no se produce. Suponiendo la utilidad proporcional al dinero, deducir la probabilidad que tal persona debe asignar, si es coherente, a una victoria del equipo visitante.

Si L representa una victoria local, X un empate y V una victoria del equipo visitante, tenemos que

$$\text{status quo} \sim \{600|L, -1.000|X \cup V\}$$

$$\text{status quo} \sim \{1.200|X, -400|L \cup V\}$$

y por lo tanto

$$0 = 600 p(L) - 1.000 \{1 - p(L)\} \rightarrow p(L) = 0,625$$

$$0 = 1.200 p(X) - 400 \{1 - p(X)\} \rightarrow p(X) = 0,250$$

de forma que

$$p(V) = 1 - p(L) - p(X) = 0,125$$

Otra forma de obtener probabilidades que describan la información de que dispone una persona que ya conoce el concepto de probabilidad es, por supuesto, preguntárselas directamente. En tal caso, puede ser conveniente proporcionar a esta persona un estímulo que le impulse a realizar un análisis cuidadoso de la información de que realmente dispone; este estímulo suele consistir en un premio cuyo valor esperado, para la persona que asigna las probabilidades, sea máximo cuando describe exactamente la información que posee.

Supongamos, por ejemplo, que, en las condiciones H , se trata de asignar probabilidades (p_1, p_2, \dots, p_k) a los sucesos inciertos mutuamente excluyentes $(\theta_1, \theta_2, \dots, \theta_k)$, de forma que $0 \leq p_i \leq 1$ y $\sum p_i = 1$. Llamémoslos θ^* al suceso incierto que finalmente tuvo lugar. Nos proponemos determinar una función de pago u , de forma que $u\{(p_1, p_2, \dots, p_k), \theta^*\}$ determine el premio obtenido por la persona que asignó las probabilidades, una vez observado el suceso θ^* que ha tenido finalmente lugar (*).

Obviamente, la función de pago u debe ser definida de forma que

- (i) estimule a mejorar la información, exigiendo que el premio obtenido sea una función creciente de la probabilidad p^* asignada al suceso θ^* que finalmente tuvo lugar.
 (ii) estimule a decir la verdad, exigiendo que el premio esperado sea máximo cuando la distribución expresada describa fielmente la información de que se dispone.

Good (1971) ha puesto de manifiesto que todas las funciones de la familia

$$u\{(p_1, \dots, p_k), \theta^*\} = \frac{A}{\alpha - 1} [\{p^*/(\sum p_i^\alpha)\}^{1/\alpha} - 1] + B \quad (4)$$

con $\alpha > 1$, $A > 0$ y B arbitraria verifican tales condiciones.

(*) Esta función de pago es un ejemplo sofisticado de función de utilidad. En la Sección 7.3 se describen algunos procedimientos que facilitan la especificación de utilidades.

En efecto, (4) es obviamente una función creciente de p^* , la probabilidad asignada al suceso que finalmente tuvo lugar.

Además, derivando parcialmente respecto a las p_i , igualando a cero y resolviendo el correspondiente sistema de ecuaciones se comprueba que el valor esperado del premio, esto es

$$\sum_{i=1}^k \left\{ \frac{A}{\alpha-1} \left[\left\{ \frac{p_i}{\sum p_i} \right\}^{\alpha} - 1 \right] + B \right\} p(\theta_i|H)$$

es máximo si, y solamente si, $p_i = p(\theta_i|H)$, donde las $p(\theta_i|H)$ representan las verdaderas opiniones del decisor.

Ejemplo 3.6.2 Calificación de exámenes.

Para obtener la máxima información posible de una pregunta con respuesta múltiple debe exigirse que se asigne una distribución de probabilidad sobre las posibles respuestas que describa los conocimientos del alumno, en lugar de limitarse a pedir que se señale la respuesta que parece más apropiada. Así, si la pregunta tiene k respuestas propuestas $\theta_1, \dots, \theta_k$ de las que se sabe que una, y solamente una, es correcta, el alumno debe asignar probabilidades $p_i = p(\theta_i|H)$, donde H son sus conocimientos, tales que $0 \leq p_i \leq 1$ y $\sum p_i = 1$.

Para calificar tales contestaciones basta, de acuerdo con la ecuación (4), con asignar

$$\frac{A}{\alpha-1} \left[\left\{ \frac{p^*}{\sum p_i} \right\}^{\alpha} - 1 \right] + B \quad (5)$$

puntos, donde $p_i = p(\theta_i|H)$ y $p^* = p(\theta^*|H)$, a una pregunta cuya contestación correcta era θ^* . Las constantes A , B y α sirven para elegir la escala de la puntuación.

Naturalmente, se obtendrá una puntuación máxima si se asigna probabilidad uno a la respuesta correcta (y por lo tanto cero a las demás) y mínima si se asigna probabilidad uno a una respuesta incorrecta (afirmando por lo tanto, de forma categórica, un error). Se pretende determinar las constantes A , B y α de forma que, con k respuestas propuestas

(i) se obtengan M puntos por una respuesta correcta (información perfecta),

- (ii) se obtengan m_0 puntos con una distribución uniforme ($1/k, \dots, 1/k$) (información nula),
- (iii) se pierdan m puntos por la afirmación categórica de un error (falsa información).

El valor de (5) cuando $p^* = 1$ (y por tanto el resto de las p_i son cero) es simplemente B . En consecuencia, (i) implica $B = M$.

El valor de (5) cuando existe una $p_i = 1$ distinta de p^* (y por lo tanto las demás probabilidades son nulas) es $B - A/(\alpha-1)$. En consecuencia, (iii) implica que $A = (\alpha-1)(M+m)$.

Finalmente, el valor de (5) cuando todas las probabilidades valen $1/k$ es

$$\frac{A}{\alpha-1} \left\{ \left(\frac{1}{k} \right)^{\alpha} - 1 \right\} + B$$

y por lo tanto (ii) implica que

$$(M+m) \left\{ \left(\frac{1}{k} \right)^{\alpha} - 1 \right\} + M = m_0$$

esto es,

$$\alpha = \log(k) / \left\{ \log(k) + \log \left(\frac{m+m_0}{M+m} \right) \right\}$$

Si, en particular, $M=2$, $m=1$, $m_0=0.25$ y $k=4$ resulta $A=5.141$, $B=2$ y $\alpha=2.714$ con lo que la calificación de una cuestión en estas condiciones, viene dada por

$$3 \left[\left\{ \frac{p^*}{\sum p_i} \right\}^{\alpha} - 1 \right] + 2 \quad \text{con } \alpha = 2.714 \quad (6)$$

donde p^* es la probabilidad asignada a la respuesta correcta.

La tabla de la página siguiente proporciona algunos valores numéricos de la expresión (6) para distintas asignaciones de probabilidad suponiendo, sin pérdida de generalidad, que la respuesta correcta era la primera. Observando (6); resulta obvio que una permutación de las probabilidades asignadas a las respuestas incorrectas no altera la puntuación obtenida.

Cómo queríamos, la puntuación oscila de 2 a -1 , decrece con la probabilidad asignada a la respuesta verdadera, y es 0.25 cuando no se sabe mostrar preferencia alguna entre las cuatro respuestas posibles. Puede observarse además que, para un mismo valor de p , la puntuación es ligeramente más alta cuando el resto de la unidad de probabilidad se distribuye de forma equilibrada entre las respuestas equivocadas; esto resulta razonable cuando se piensa que asignar una probabilidad alta a una respuesta equivocada es mostrarse convencido de un error. Finalmente, es obvio a la vista de (6) que si se asigna probabilidad cero a la respuesta correcta la puntuación es -1 cualesquiera que sean las demás probabilidades asignadas.

3.7. Discusión y referencias

Las principales propiedades matemáticas de la probabilidad se conocen hace siglos, pero no fueron rigurosamente establecidas hasta que Kolmogo-

$p_1 = p^*$	p_2	p_3	p_4	u
1.00	0.00	0.00	0.00	2.000
0.90	0.10	0.00	0.00	1.995
0.80	0.10	0.10	0.00	1.987
0.80	0.20	0.00	0.00	1.957
0.70	0.10	0.10	0.10	1.971
0.70	0.20	0.10	0.00	1.929
0.70	0.30	0.00	0.00	1.824
0.60	0.20	0.10	0.10	1.881
0.60	0.20	0.20	0.00	1.822
0.60	0.30	0.10	0.00	1.731
0.60	0.40	0.00	0.00	1.502
0.50	0.20	0.20	0.10	1.704
0.50	0.30	0.20	0.00	1.502
0.50	0.40	0.10	0.00	1.267
0.50	0.50	0.00	0.00	0.936
0.40	0.20	0.20	0.20	1.365
0.40	0.30	0.20	0.10	1.200
0.40	0.40	0.10	0.10	0.909
0.40	0.50	0.10	0.00	0.546
0.40	0.60	0.00	0.00	0.249
0.30	0.30	0.20	0.20	0.615
0.30	0.30	0.30	0.10	0.483
0.30	0.40	0.20	0.10	0.344
0.30	0.40	0.30	0.00	0.215
0.30	0.50	0.20	0.00	0.042
0.30	0.50	0.10	0.10	0.072
0.30	0.60	0.10	0.00	-0.167
0.25	0.25	0.25	0.25	0.250
0.20	0.30	0.30	0.20	-0.194
0.20	0.80	0.00	0.00	-0.725
0.10	0.30	0.30	0.30	-0.774
0.10	0.90	0.00	0.00	-0.931
0.00	0.40	0.30	0.30	-1.000
0.00	1.00	0.00	0.00	-1.000

rov (1933) propuso una definición axiomática basada en las propiedades de las frecuencias relativas que considera a la probabilidad como un caso particular de medida finita en que el espacio total mide la unidad. El hecho notable de que los grados de creencia se comporten de acuerdo con las mismas leyes que las frecuencias relativas (Teorema 2.4.1) da lugar a una teoría unificada de la probabilidad, independiente de la interpretación que se le atribuya.

La mayor parte de los textos de teoría de la probabilidad adoptan la definición axiomática de probabilidad de Kolmogorov, basada en probabilidades «absolutas», a la que añaden una definición de probabilidad condicionada. Puesto que todas las probabilidades son condicionadas (aunque a menudo se omitan las condiciones para abreviar la notación) parece más oportuno desarrollar la teoría de la probabilidad partiendo directamente de las propiedades

de las probabilidades condicionadas. Estas propiedades (las del Teorema 2.4.1) fueron tomadas por Renyi (1962/1966) como base de su definición axiomática. El texto de Lindley (1965) se basa en esta misma definición.

Los libros de Feller (1957/1966) y Gnedenko (1962) se ocupan de las propiedades matemáticas de la probabilidad a un nivel relativamente elemental. Con respecto a la interpretación de la probabilidad, son importantes los trabajos contenidos en Kyburg & Smokler (1964) y los de Laplace (1812/1912), Keynes (1921), Jeffreys (1939/1967), Reichenbach (1969), Braithwaite (1953), Carnap (1950), Good (1950), De Finetti (1970/1975) y Fine (1973).

No nos hemos ocupado apenas del lado *combinatorio* de la probabilidad cuyas aplicaciones se centran casi exclusivamente en los juegos de azar. Casi todos los textos de probabilidad incluyen una discusión elemental de combinatoria; puede consultarse, por ejemplo, Feller (1957/1968) o a un nivel más elemental DeGroot (1975). Los libros de Whitworth (1901) y David & Barton (1962) tratan estos problemas de forma monográfica.

El postulado de σ -aditividad es aceptado de forma casi universal debido a la simplificación matemática a que, según veremos, da lugar. DeGroot (1970) da un razonamiento sobre su plausibilidad. Una notable excepción es De Finetti (1970/1975) quien desarrolla su *Teoría de la Probabilidad* sin utilizarlo.

En nuestro tratamiento del Teorema de Bayes, hemos subrayado su utilización sistemática para incorporar información adicional. Esta es una característica específica de la escuela *Bayesiana* de inferencia dentro de cuyos presupuestos está escrito este libro. Conviene subrayar sin embargo, que el Teorema de Bayes, como tal Teorema, es un resultado matemático que se deduce indiscutiblemente de las propiedades de la medida de probabilidad; es la interpretación de las probabilidades como grados de creencia, y no el contenido formal del Teorema de Bayes lo que se haya sujeto a controversia.

Existe una importante bibliografía sobre el tema de la asignación de probabilidades. La referencia clave es Savage (1971); dos trabajos recientes sobre este tema son los de Murphy & Winkler (1975) y Spetzler & Staël von Holstein (1975). Son importantes los comentarios de Lindley (1978) sobre la conveniencia de «extender la conversación» y estimar varias probabilidades relacionadas entre sí por las leyes de la probabilidad como, por ejemplo, $p(A)$, $p(B)$, $p(A|B)$ y $p(A|\bar{B})$ que deben cumplir $p(A) = p(A|B)p(B) + p(A|\bar{B})\{1 - p(B)\}$; en efecto, esto permite verificar la coherencia del decisor y mejorar la precisión de sus estimaciones. La idea ha sido recogida en Bernardo & Basulto (1978).

PROBLEMAS

1. A las elecciones para la alcaldía de una ciudad se presentan únicamente tres candidatos A , B y C . Si se juzga que $p(A) = p(B) = 0,4$, construir el espacio probabilístico relevante a un problema de decisión cuyas consecuencias dependen del candidato elegido.
2. Se sabe que el 1% de los individuos de una población son daltónicos. Determinar el tamaño mínimo n de una muestra de esa población para que la probabilidad de que contenga al menos un daltónico sea como mínimo 0,95.
3. Determinar la probabilidad de obtener 8 unos en 10 lanzamientos de dado. Repetir el problema suponiendo que se asigna probabilidad $1/4$ a que el dado esté cargado de forma que las probabilidades de obtener 1, 2, 3, 4, 5 o 6 sean respectivamente $1/5$, $1/6$, $1/6$, $1/6$, $1/6$, $2/15$.
4. Una pareja ha tenido tres hijos varones consecutivamente. Se oye a menudo argumentar que es más probable que el cuarto nacimiento sea el de una niña puesto que se sabe que, aproximadamente, hay el mismo número de nacimientos de cada sexo. Comentar la falacia del argumento.
5. Se sabe que entre los 120 estudiantes de una residencia, hay 60 que estudian Medicina, 50 que estudian Farmacia y 20 que cursan ambos estudios simultáneamente. Determinar la probabilidad de que uno de ellos, escogido al azar, estudie Medicina o Farmacia y la de que *no estudie* ambas simultáneamente. Calcular, además, la probabilidad de que estudie Medicina un alumno del que va se sabe que estudia Farmacia.
6. Tenemos tres cajas de inyectables; la caja A contiene 10 unidades, de las cuales 4 están alteradas; la caja B , 6 con una alterada y la caja C , 8 con tres alteradas. Si escogemos una caja al azar, y extraemos un inyectable alterado. ¿Cuál es la probabilidad de que tal inyectable perteneciese a la caja B ?
7. Se dice que en las condiciones H la ocurrencia de B favorece la de A si $p(A|B, H) > p(A|H)$. Supongamos que, en las mismas condiciones, la ocurrencia de A favorece la de B y la ocurrencia de B favorece la de C . ¿Es cierto entonces que la ocurrencia de A favorece la de C ? Demostrarlo en caso afirmativo o dar un contraejemplo en caso negativo.
8. Un médico cree que su paciente tiene una de tres enfermedades A , B , C (cada una de las cuales es incompatible con las otras dos) cuyas probabilidades son, respectivamente,

$$p(A) = 0,2 \quad p(B) = 0,4 \quad p(C) = 0,4$$

Las tres enfermedades exigen una intervención quirúrgica para su curación. La probabilidad de éxito de dicha intervención depende de la enfermedad del paciente, y su estimación por el médico es la siguiente:

$$p(\text{Éxito}|A) = 0,5 \quad p(\text{Éxito}|B) = 0,8 \quad p(\text{Éxito}|C) = 0,7$$

¿Cuál es la probabilidad de éxito para la intervención quirúrgica?, ¿y la de fracaso?

9. Un test médico diseñado para la diagnosis de una enfermedad detecta dicha enfermedad en el 90% de los pacientes que la padecen y que son sometidos al mismo, pero da también positivo en el 2% de personas que sin padecerla son sometidos al test. Sin embargo, al pasar el test a una población determinada, el porcentaje de contestaciones nada menos que positivas del test correspondientes a personas *no* afectadas por la enfermedad es del 50%. Explicar este hecho en función de la proporción de personas no afectadas por la enfermedad en la población estudiada.
10. Un paciente que piensa que puede tener cáncer consulta a un médico A , del que sabe que diagnostica cáncer únicamente al 60% de los que lo tienen y nunca a los que no lo tienen. El médico A no le diagnostica cáncer pero, para estar más seguro, consulta a otro médico B del que sabe que diagnostica cáncer al 80% de los que lo tienen y al 10% de los que no lo tienen. ¿Qué probabilidad inicial de tener cáncer debe tener el paciente para que la probabilidad final, después de que ni A ni B le hayan diagnosticado cáncer sea todavía 0,5? ¿Y para que sea 0,1?

Cantidades aleatorias

En este capítulo, se introduce el concepto de *cantidad aleatoria* para poder traducir el espacio probabilístico relevante a un contexto numérico más fácilmente manipulable.

Se define la *función de distribución* de una cantidad aleatoria, poniendo de manifiesto que contiene toda la información que puede necesitarse sobre ella. Se distingue entre cantidades aleatorias discretas y continuas, y se introducen las *densidades* de probabilidad, insistiendo en su interpretación intuitiva. Se estudia el concepto de *función* de una cantidad aleatoria.

Se definen y estudian los *momentos* de una distribución de probabilidad, poniendo de manifiesto la forma en que permiten una descripción aproximada de la distribución. Se introducen las *funciones generatrices* subrayando su doble utilidad para identificar una distribución y para obtener sus momentos.

Se introducen los *vectores aleatorios* o cantidades aleatorias multidimensionales, se definen las correspondientes distribuciones marginales y condicionales, y se estudian las *funciones* de vectores aleatorios.

Resulta conveniente poder representar de forma numérica el espacio probabilístico que describe una determinada situación incierta. Utilizaremos para ello una *función* que asocie un número real a cada uno de los elementos del conjunto de referencia. A esta función se le suele llamar *variable aleatoria*, algo inaudito tratándose de una función. De Finetti propuso llamarle *cantidad aleatoria*, sugerencia que nosotros hemos aceptado.

4.1. Cantidad aleatoria y función de distribución

En algunas situaciones, los sucesos inciertos relevantes en un problema de decisión pueden expresarse directamente en forma numérica; así, por ejemplo, aunque el estado de salud de un paciente es difícil de medir directamente podemos cuantificar distintas características suyas: temperatura, presión sanguínea, cantidad de hematies en la sangre, etc.

En otras situaciones, sin embargo, los sucesos inciertos relevantes no son de tipo numérico; así, los efectos de un tratamiento hormonal dependen de que el paciente sea hombre o mujer y la incidencia de una enfermedad depende de la región geográfica considerada; sin embargo, con objeto de poder trabajar siempre en términos numéricos se procura, en estos casos, *asignar* un número a cada uno de los posibles sucesos inciertos. Así, podríamos asignar el número 0 a los hombres y el 1 a las mujeres, y podríamos asignar a cada región geográfica su extensión, su altura máxima, o simplemente su número por orden alfabético. Una *cantidad aleatoria* es una función que asigna un número a cada suceso incierto. Más formalmente,

DEFINICIÓN 4.1.1. Dado un espacio probabilístico (Ω, Σ, P) , una cantidad aleatoria es una función que asocia un número real a cada elemento de Ω .

En realidad, no todas las funciones sirven como cantidades aleatorias. En efecto, los sucesos en que estamos interesados en la recta real son los intervalos o sus combinaciones. Queremos, pues, definir la función de forma que la imagen inversa de cualquier intervalo real sea un suceso perteneciente al álgebra de sucesos interesantes, con objeto de poder determinar su probabilidad. Se dice entonces que la función considerada es *medible*. Una cantidad aleatoria debe ser una función medible; todas las funciones que se utilizan en este libro lo son.

Naturalmente, unos valores de la cantidad aleatoria resultarán más probables que otros porque corresponden a sucesos más probables. Así, dado un espacio probabilístico (Ω, Σ, P_1) , la medida de probabilidad P_1 definida sobre todos los elementos de Σ dará lugar a una medida de probabilidad P_2 sobre los sucesos interesantes de la recta real R , esto es, los intervalos y sus combinaciones, de forma que si X es una cantidad aleatoria

$$P_2\{X \in I\} = P_1\{\omega; X(\omega) \in I\}$$

Así la probabilidad de que la cantidad aleatoria X tome un valor perteneciente al intervalo I es simplemente la probabilidad de que tenga lugar uno de los sucesos inciertos ω cuya imagen, mediante la función X , pertenece a I .

Ejemplo 4.1.1. Temperaturas

En un modelo simplificado, el estado general de un determinado tipo de paciente debe ser uno de los elementos del conjunto $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7\}$. Se sabe que a ω_1 le corresponden temperaturas inferiores a las normales, a ω_2 y a ω_3 temperaturas normales, a ω_4 y ω_5 fiebre moderada y a ω_6 y ω_7 fiebre alta y muy alta respectivamente. Definir una cantidad aleatoria que a cada estado general le haga corresponder su temperatura media; suponiendo que las probabilidades de los distintos estados generales son

$$p(\omega_1) = p(\omega_7) = 0,05 \quad p(\omega_2) = 0,25 \quad p(\omega_3) = 0,35 \\ p(\omega_4) = p(\omega_5) = p(\omega_6) = 0,1$$

determinar la probabilidad de que tal cantidad aleatoria tome un valor mayor de 39 °C, y la que lo tome entre 36 y 37 °C.

De la experiencia adquirida con otros pacientes, se sabe que las temperaturas medias correspondientes a los tipos de fiebre descritos son, aproximadamente, 35,6, 36,6, 37,7, 39,9 y 40,6 °C respectivamente. En consecuencia, la cantidad aleatoria deseada X quedará definida mediante

$$X(\omega_1) = 35,6 \\ X(\omega_2) = X(\omega_3) = 36,6 \\ X(\omega_4) = X(\omega_5) = 37,7 \\ X(\omega_6) = 39,9 \\ X(\omega_7) = 40,6$$

y, por lo tanto,

$$P[X > 39] = P[X = 40,6 \cup X = 39,9] = P[\omega_6] + P[\omega_7] = 0,1 + 0,05 = 0,15 \\ P[36 < X < 37] = P[X = 36,6] = P[\omega_2] + p[\omega_3] = 0,25 + 0,35 = 0,60$$

Todas las probabilidades relativas a los valores que puede tomar una cantidad aleatoria pueden obtenerse directamente a partir de una función, la *función de distribución* que representa así un útil resumen de los valores que la variable puede tomar y de las probabilidades con que los toma.

DEFINICIÓN 4.1.2. La función de distribución de una cantidad aleatoria X , que representaremos con la letra F es la función definida mediante

$$F(x) = P[X \leq x], \quad -\infty < x < \infty$$

Así, el valor de la función de distribución de una cantidad aleatoria en el punto real x es igual a la probabilidad de que la cantidad aleatoria X tome un valor menor o igual a x .

TEOREMA 4.1.1. Toda función de distribución F de una cantidad aleatoria satisface las propiedades siguientes:

- (i) *No decreciente:* $x_1 < x_2 \rightarrow F(x_1) \leq F(x_2)$
- (ii) $\lim_{x \rightarrow \infty} F(x) = 1$ y $\lim_{x \rightarrow -\infty} F(x) = 0$

Demostración

En efecto, puesto que

$$F(x_1) = P\{]-\infty, x_1]\}, F(x_2) = P\{]-\infty, x_2]\}, x_1 < x_2$$

tenemos que $]-\infty, x_1] \subset]-\infty, x_2]$ y por lo tanto, en virtud del Teorema 3.2.3 (ii) $P\{]-\infty, x_1]\} \leq P\{]-\infty, x_2]\}$, lo que demuestra (i)

Si $x \rightarrow \infty$, $F(x)$ tiende a $P\{]-\infty, \infty[\} = P\{\Omega\} = 1$ y si $x \rightarrow -\infty$, $F(x)$ tiende a $P\{\emptyset\} = 0$, lo que prueba (ii).

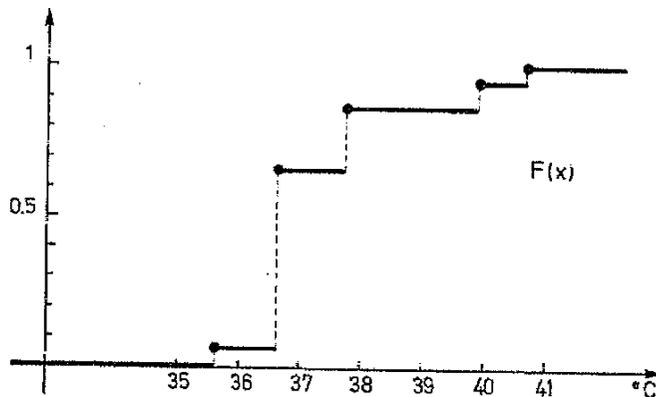
Ejemplo 4.1.2. Temperaturas (cont.)

Construir y representar la función de distribución correspondiente a la cantidad aleatoria definida en el Ejemplo 4.1.1.

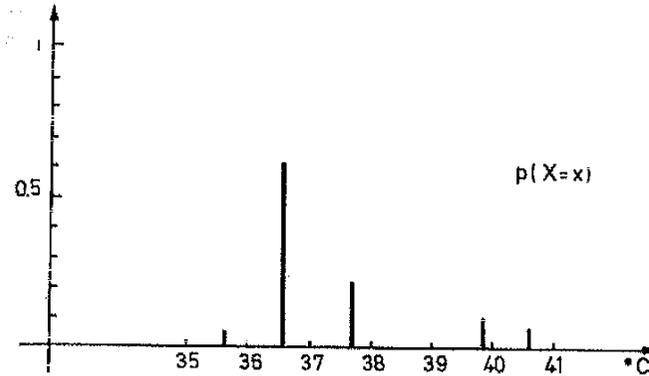
Por definición $F(x) = P\{X \in]-\infty, x]\}$ y por tanto,

- $-\infty < x < 35,6 \rightarrow F(x) = P\{\emptyset\} = 0$
- $35,6 \leq x < 36,6 \rightarrow F(x) = P\{\omega_1\} = 0,05$
- $36,6 \leq x < 37,7 \rightarrow F(x) = P\{\omega_1 \cup \omega_2 \cup \omega_3\} = 0,65$
- $37,7 \leq x < 39,9 \rightarrow F(x) = P\{\omega_1 \cup \omega_2 \cup \omega_3 \cup \omega_4 \cup \omega_5\} = 0,85$
- $39,9 \leq x < 40,6 \rightarrow F(x) = P\{\omega_1 \cup \omega_2 \cup \omega_3 \cup \omega_4 \cup \omega_5 \cup \omega_6\} = 0,95$
- $40,6 \leq x \leq +\infty \rightarrow F(x) = P\{\Omega\} = 1$

La correspondiente representación gráfica es una función en escalera:



Los valores posibles de la variable aleatoria X son 35,6, 36,6, 37,7, 39,9 y 40,6, que se obtienen respectivamente con probabilidades 0,05, 0,6, 0,2, 0,1, 0,05, situación que puede describirse mediante la gráfica



Observese como los valores de las probabilidades $p(X = x_i)$, corresponden exactamente a los «saltos» de la función de distribución $F(x)$.

4.2. Distribuciones discretas

Las cantidades aleatorias pueden ser esencialmente de dos tipos diferentes, según el número de valores distintos que pueden tomar.

DEFINICIÓN 4.2.1. Se dice que una cantidad aleatoria X es discreta si solo puede tomar un número finito, o infinito numerable (*), de valores distintos. Su función de probabilidad $p(x)$ se define mediante

$$p(x) = p\{X = x\}$$

La cantidad aleatoria definida en el Ejemplo 4.1.1. era discreta. Para todo número x que no sea uno de los valores posibles de la cantidad aleatoria X tendremos, evidentemente, $p(x) = 0$. Por otro lado, si la sucesión $\{x_1, x_2, \dots\}$ incluye todos los valores posibles de X tendremos $\sum p(x_i) = 1$. Así, una cantidad aleatoria discreta da lugar a una *distribución discreta* de probabilidad, entre el conjunto, finito o numerable, de sus valores posibles, de forma que $0 < p(x_i) < 1$ y $\sum p(x_i) = 1$. Naturalmente, distintas cantidades aleatorias pueden dar lugar a la misma distribución de probabilidad.

La función de probabilidad de una cantidad aleatoria discreta, esto es, la

(*) Un conjunto A es *infinito numerable* si es posible establecer una biyección entre A y los números naturales.

función que a cada número real le asocia su probabilidad, suele representarse gráficamente mediante una colección de segmentos verticales situados en cada uno de los valores posibles y con una longitud proporcional a la de su probabilidad. La función de distribución de una cantidad aleatoria discreta es una función en escalera con saltos iguales a $p(x_i)$ en cada uno de los valores x_i que puede tomar (ver Ejemplo 4.1.2.).

Frecuentemente, la tarea de determinar la función de probabilidad de una cantidad aleatoria discreta se reduce a la de resolver un sencillo problema de probabilidades.

Ejemplo 4.2.1. Elección de pacientes

En la sala de espera de una policlínica se encuentran en cola 3 hombres y 7 mujeres, de los que se hace pasar a los 2 primeros. Determinar la función de probabilidad del número de hombres que pueden entrar.

La correspondiente cantidad aleatoria X puede tomar los valores 0, 1, y 2. Las probabilidades respectivas serán

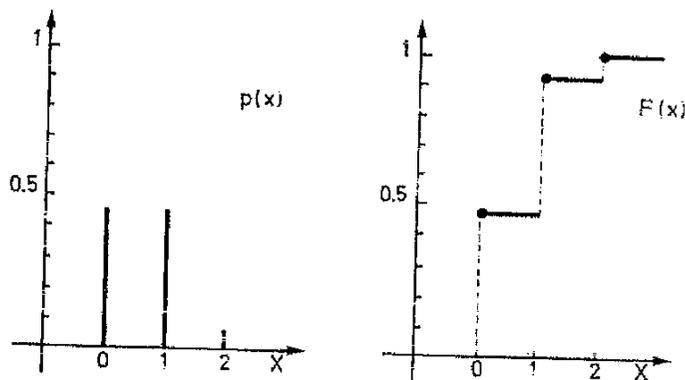
$$p(x=0) = \frac{7}{10} \cdot \frac{6}{9} = \frac{7}{15}$$

$$p(x=1) = \frac{3}{10} \cdot \frac{7}{9} + \frac{7}{10} \cdot \frac{3}{9} = \frac{7}{15}$$

$$p(x=2) = \frac{3}{10} \cdot \frac{2}{9} = \frac{1}{15}$$

La función de probabilidad de la cantidad aleatoria X toma pues los valores $7/15$, $7/15$, $1/15$, en los puntos 0, 1 y 2.

Las representaciones gráficas de esta función y de su correspondiente función de distribución son



La más sencilla de las distribuciones discretas es la distribución de Bernoulli.

DEFINICIÓN 4.2.2. Una cantidad aleatoria discreta X tiene una distribución de Bernoulli con parámetro θ , si puede tomar los valores 1 y 0, y lo hace con probabilidades θ y $1 - \theta$. Consecuentemente, su función de probabilidad, que denotaremos $Br(x|\theta)$ será de la forma

$$p(X=x) = Br(x|\theta) = \theta^x(1-\theta)^{1-x} \quad \text{para } x = 0, 1$$

$$= 0, \quad \text{para cualquier otro valor de } x$$

donde $0 < \theta < 1$.

Entre las distribuciones discretas de probabilidad más utilizadas se encuentra la *distribución binomial*, íntimamente relacionada con la distribución de Bernoulli, y con las sucesiones del mismo nombre, estudiadas en la Sección 3.3.

DEFINICIÓN 4.2.3. Una cantidad aleatoria discreta X tiene una distribución binomial si su función de probabilidad, que denotaremos $Bi(x|\theta, n)$, es de la forma

$$p(X=x) = Bi(x|\theta, n) = \binom{n}{x} \theta^x(1-\theta)^{n-x}, \quad \text{para } x = 0, 1, \dots, n$$

$$= 0, \quad \text{para cualquier otro valor de } x$$

donde $0 < \theta < 1$.

Consideremos una sucesión de Bernoulli $\{A_i, i = 1, 2, \dots\}$. Supongamos $p(A_i) = \theta$ y consideremos n sucesos cualesquiera de la sucesión; entonces, la probabilidad de que ocurran exactamente x de ellos, $0 \leq x \leq n$ es, según vimos en la Sección 3.3, $\binom{n}{x} \theta^x(1-\theta)^{n-x}$. En consecuencia, el número de ocurrencias en una sucesión de Bernoulli es una cantidad aleatoria con una distribución Binomial cuyos parámetros, θ y n , son respectivamente la probabilidad de que ocurra uno de los sucesos y el número de sucesos considerado.

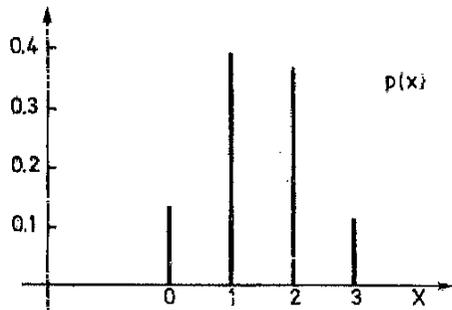
Ejemplo 4.2.2. Sexo de un recién nacido (cont.)

Representar la función de probabilidad y la función de distribución de la cantidad aleatoria X que describe el número de niñas que pueden nacer en los tres próximos alumbramientos que tengan lugar en el Hospital Clínico.

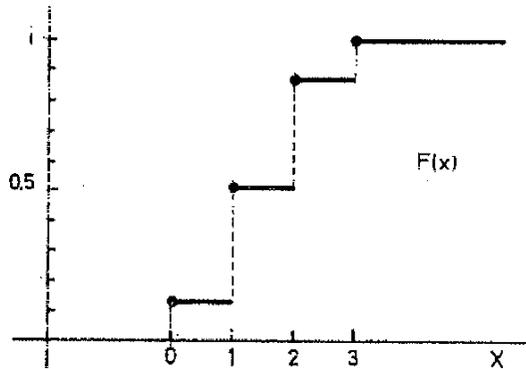
Como vimos en el Ejemplo 3.3.2, las probabilidades correspondientes resultan ser

$$\begin{aligned}
 p(X = 0) &= B(0|0.49, 3) = 0.1237 \\
 p(X = 1) &= B(1|0.49, 3) = 0.3823 \\
 p(X = 2) &= B(2|0.49, 3) = 0.3675 \\
 p(X = 3) &= B(3|0.49, 3) = 0.1175
 \end{aligned}$$

y, en consecuencia, la función de probabilidad es



y la correspondiente función de distribución



En ocasiones, se manejan cantidades aleatorias cuya distribución es Binomial con θ muy pequeño y n muy grande. En estas condiciones, el cálculo de la correspondiente función de probabilidad $B(X|\theta, n)$ resulta muy laborioso, pero puede obtenerse fácilmente una aproximación adecuada en función del producto $\lambda = \theta n$. En efecto,

TEOREMA 4.2.1. Para $n \rightarrow \infty$ y $\theta \rightarrow 0$, con $n\theta = \lambda$,
 $\lim_{n \rightarrow \infty} B(k|\theta, n) = e^{-\lambda} \lambda^k / k!$

Demostración

Utilizando la definición de número combinatorio (Ecuación 3.2.3),

$$\begin{aligned}
 B(k|\theta, n) &= \frac{n(n-1) \dots (n-k+1)}{k!} \theta^k (1-\theta)^{n-k} \\
 &= \frac{\lambda^k}{k!} \frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}
 \end{aligned}$$

y tomando límites, cuando $n \rightarrow \infty$

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \frac{n}{n} \frac{n-1}{n} \dots \frac{n-k+1}{n} \left(1 - \frac{\lambda}{n}\right)^{-k} &= 1 \\
 \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n &= e^{-\lambda}
 \end{aligned}$$

resulta $\lim_{n \rightarrow \infty} B(k|\theta, n) = e^{-\lambda} \lambda^k / k!$

El resultado anterior sugiere la definición de una nueva distribución de probabilidad.

DEFINICIÓN 4.2.4. Una cantidad aleatoria discreta X tiene una distribución de Poisson si su función de probabilidad, que denotaremos $Po(x|\lambda)$, es de la forma

$$\begin{aligned}
 p(X = x) &= Po(x|\lambda) = e^{-\lambda} \lambda^x / x!, \quad x = 0, 1, 2, \dots \\
 &= 0 \text{ en cualquier otro caso}
 \end{aligned}$$

donde $\lambda > 0$.

De esta forma, el Teorema 4.2.1 puede enunciarse de la forma siguiente

$$\lim_{n \rightarrow \infty, \theta \rightarrow 0} B(x|n, \theta) = Po(x|\lambda) \tag{1}$$

Ejemplo 4.2.3. Probabilidad de cáncer

Se sabe que entre las personas sometidas a la acción de un cierto agente cancerígeno, el 2 % llegan a contraer la enfermedad. Determinar la probabilidad de que más de una persona contraiga cáncer en un grupo de sesenta que se han visto sometidas a la acción del mencionado agente.

La cantidad aleatoria X que describe el número de personas que contraerán la enfer-

medad tiene obviamente una distribución binomial $Bi(x|0,02, 60)$, con $\theta = 0,02$, pequeño, y $n = 60$, grande. En virtud del Teorema 4.2.1, tal distribución puede aproximarse por una distribución de Poisson $Po(x|1,2)$ con parámetro $\lambda = 0,02 \times 60 = 1,2$. En consecuencia, la probabilidad de que exactamente k personas de las 60 contraigan la enfermedad es aproximadamente

$$P[X = k] = Po(k|1,2) = e^{-1,2}(1,2)^k/k!$$

y por tanto la probabilidad pedida será

$$P[X > 1] = 1 - P[X = 0] - P[X = 1] = 1 - (1 + 1,2)/e^{1,2} = 0,3374$$

4.3. Distribuciones continuas

Frecuentemente, las cantidades aleatorias pueden tomar un número infinito, no numerable, de valores distintos. Gracias al postulado de σ -aditividad, la probabilidad de que la variable tome un valor perteneciente a un determinado intervalo puede expresarse mediante una integral (*).

DEFINICIÓN 4.3.1. Se dice que una cantidad aleatoria X es continua si existe una función no-negativa $p(x)$, llamada función de densidad de probabilidad tal que para todo intervalo (a, b) de la recta real R ,

$$P[X \in (a, b)] = \int_a^b p(x)dx$$

Una notación más precisa para la función de densidad y la función de distribución de una cantidad aleatoria X , que emplearemos cuando resulte necesario, consiste en utilizar, respectivamente $p_X(x)$ y $F_X(x)$. Generalmente, sin embargo, puede utilizarse sin ambigüedad la notación simplificada $p(x)$ y $F(x)$.

La cantidad aleatoria que al último paciente ingresado en el hospital le asocia su altura es, conceptualmente, una variable continua. En la práctica, debido a los errores de medida, las cantidades aleatorias son casi siempre discretas (la altura se registra generalmente como un número entero de centímetros, con lo que 160,333... no resulta un valor «posible» de la cantidad aleatoria). Sin embargo resulta conveniente, puesto que simplifica enormemente los cálculos sin modificar apreciablemente el resultado, describir como continuas aquellas cantidades aleatorias que conceptualmente lo son.

Una cantidad aleatoria continua da lugar a una *distribución continua* de probabilidad, esto es, a una *distribución* de la unidad de probabilidad de forma

(*) Se trata de un resultado de teoría de la medida, cuya descripción detallada no es posible al nivel matemático escogido para este libro.

continua entre el conjunto de sus valores posibles con mayor *densidad* en aquellas zonas más verosímiles, esto es, en aquellos valores para los que la función de densidad de probabilidad $p(x)$ es mayor.

La función de densidad *no* describe probabilidades sino densidades de probabilidad. Sin embargo, de acuerdo con la Definición 4.3.1 la probabilidad de cualquier *intervalo* puede ser determinada por integración a partir de la función de densidad de probabilidad.

La probabilidad de que una cantidad aleatoria *continua* X tome un valor determinado x cualquiera es *cero*. En efecto, por definición,

$$P[X = x] = \int_x^x p(x)dx = 0$$

lo que subraya el hecho de que *no* todos los sucesos de probabilidad cero son *imposibles*. En este caso, el suceso $X = x$ es perfectamente posible, pero su probabilidad es cero porque su «medida», la de un punto, es infinitamente pequeña comparada con la «medida» de un intervalo.

TEOREMA 4.3.1. Toda función de densidad de probabilidad satisface las propiedades siguientes

$$(i) \int_{-\infty}^{\infty} p(x)dx = 1$$

$$(ii) \text{ En los puntos de continuidad de } F(x), p(x) = dF(x)/dx.$$

Demostración

Una cantidad aleatoria es una función que sólo toma valores reales; por lo tanto $P[X \in R] = 1$. Por otro lado, por definición,

$$P[X \in R] = \int_{-\infty}^{\infty} p(x)dx$$

lo que demuestra (i). Además, en virtud de las Definiciones 4.1.2 y 4.3.1,

$$F(x) = P[X \leq x] = P\{X \in]-\infty, x]\} = \int_{-\infty}^x p(x)dx$$

y derivando obtenemos (ii).

La más sencilla de las distribuciones continuas es la distribución *uniforme*.

DEFINICIÓN 4.3.2. Una cantidad aleatoria continua X tiene una distribución

uniforme en el intervalo $[\alpha, \beta]$ si su función de densidad de probabilidad, que denotaremos $Un(x|\alpha, \beta)$ es de la forma

$$p(x) = Un(x|\alpha, \beta) = \frac{1}{\beta - \alpha} \text{ para } \alpha \leq x \leq \beta$$

$$= 0, \text{ para cualquier otro } x$$

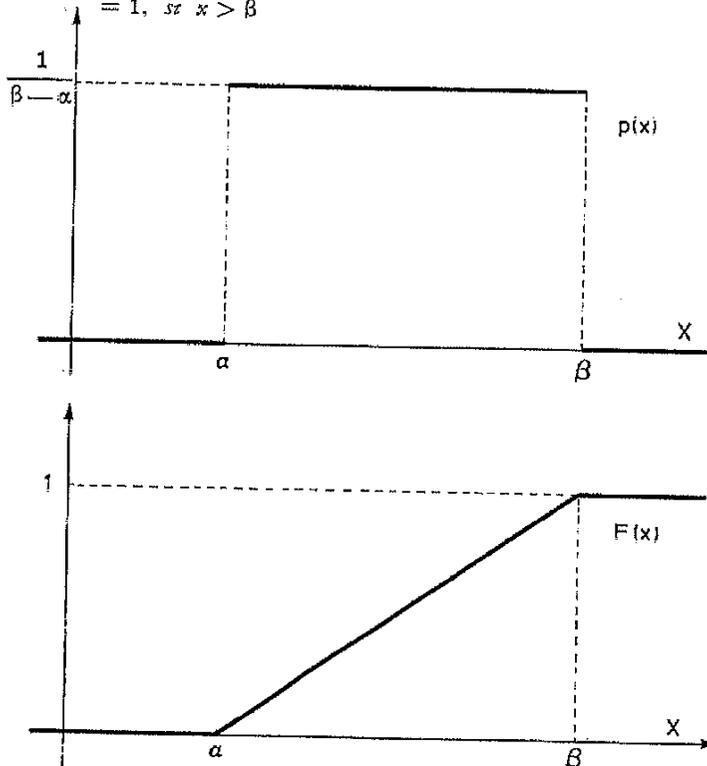
donde $\alpha \in R, \beta \in R$ y $\alpha < \beta$.

Su función de distribución es, claramente,

$$F(x) = 0, \text{ si } x < \alpha$$

$$= \int_{\alpha}^x Un(x|\alpha, \beta) dx = \frac{1}{\beta - \alpha} \int_{\alpha}^x dx = \frac{x - \alpha}{\beta - \alpha}, \text{ si } \alpha \leq x \leq \beta$$

$$= 1, \text{ si } x > \beta$$



Para definir la función de densidad de probabilidad de una cantidad aleatoria continua X basta especificar una función $f(x)$ tal que $p(x) \propto f(x)$. En efecto, si se sabe que $p(x) = C f(x)$, la constante de proporcionalidad C puede ser determinada utilizando el hecho de que una densidad de probabilidad debe integrar uno; así,

$$\int_{-\infty}^{\infty} p(x) dx = C \int_{-\infty}^{\infty} f(x) dx = 1$$

y, por lo tanto,

$$C = 1 / \int_{-\infty}^{\infty} f(x) dx$$

Esta última integral puede resultar difícil. En los ejemplos de distribuciones continuas que damos a continuación, se especifica el valor de la constante de proporcionalidad correspondiente a cada uno de ellos aunque, como veremos más adelante, no suele ser necesario conocer su valor.

Frecuentemente, se quiere describir la información de que se dispone sobre una cantidad aleatoria que solo puede tomar valores en el intervalo $]0, 1[$, por lo que resulta interesante disponer de una familia de distribuciones de probabilidad cuyas correspondientes densidades solo tomen valores no nulos en ese intervalo. La familia de distribuciones Beta da lugar, variando los valores de sus parámetros, a infinitas densidades con esas características.

DEFINICIÓN 4.3.3. Una cantidad aleatoria continua X tiene una distribución Beta si su función de densidad de probabilidad, que denotaremos $Be(x|\alpha, \beta)$ es de la forma

$$p(x) = Be(x|\alpha, \beta) = C x^{\alpha-1} (1-x)^{\beta-1}, \text{ si } 0 < x < 1$$

$$= 0, \text{ para cualquier otro } x$$

donde $\alpha > 0$ y $\beta > 0$. El valor de la constante de proporcionalidad es $C = \Gamma(\alpha + \beta) / \Gamma(\alpha)\Gamma(\beta)$ (*).

La distribución uniforme *standard*, $Un(0|0, 1)$ es un caso particular de distribución Beta, concretamente la $Be(1|1, 1)$.

(*) Por definición, la función gamma, denotada $\Gamma(x)$ es el valor de la integral definida $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$. Puede demostrarse además que, para x , $\Gamma(x+1) = x\Gamma(x)$ y $\Gamma(1) = 1$. En consecuencia, para valores enteros, $\Gamma(n) = (n-1)!$.

También resulta frecuente desear describir la información de que se dispone sobre una cantidad aleatoria que sólo puede tomar valores positivos. La familia de distribuciones *Gamma* da lugar, variando los valores de sus parámetros, a infinitas densidades de probabilidad que sólo toman valores no nulos en el intervalo $]0, \infty[$.

DEFINICIÓN 4.3.4. Una cantidad aleatoria continua X tiene una distribución *Gamma* si su función de densidad de probabilidad, que denotaremos $Ga(x|\alpha, \beta)$ es de la forma

$$p(x) = Ga(x|\alpha, \beta) = Cx^{\alpha-1}e^{-\beta x}, \text{ si } x > 0 = 0, \\ = 0, \text{ para cualquier otro } x$$

donde $\alpha > 0$ y $\beta > 0$. El valor de la constante de proporcionalidad es $C = \beta^\alpha / \Gamma(\alpha)$.

El caso particular de la distribución *Gamma* en que $\alpha = 1$ recibe el nombre de distribución *exponencial* con parámetro β , $Ex(x|\beta)$, y el caso particular que $\alpha = n/2$ y $\beta = 1/2$ el de distribución χ^2 con n grados de libertad.

En numerosas ocasiones, las propias condiciones del problema sirven para especificar la forma de la función de densidad.

Ejemplo 4.3.1. Longitud de una úlcera

Se sabe que, bajo determinadas circunstancias, la probabilidad con que se presenta un tipo de úlcera es inversamente proporcional a la raíz cuadrada de su longitud, que puede llegar a ser de 16 mm. Determinar y representar la función de densidad de probabilidad y la función de distribución de la cantidad aleatoria longitud de la úlcera. Determinar la probabilidad de que la úlcera sea mayor de 8 mm.

Por hipótesis, se sabe que

$$p(x) = C/\sqrt{x}, \quad 0 < x < 16 \\ = 0, \text{ en caso contrario}$$

y puesto que $\int_0^{16} p(x)dx = 1$, sabemos que

$$\int_0^{16} p(x)dx = C \int_0^{16} \frac{1}{\sqrt{x}} dx = 2C\sqrt{x} \Big|_0^{16} = 2C\sqrt{16} = 8C = 1$$

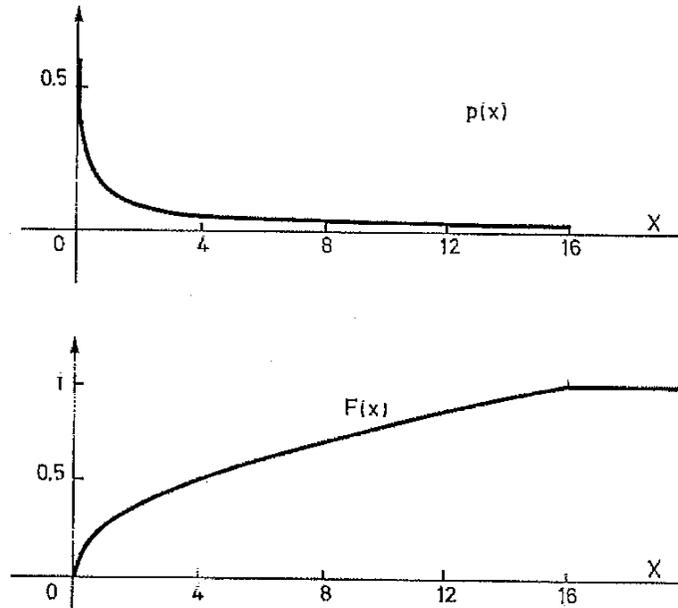
y por tanto $C = 1/8$. En consecuencia, la función de densidad de probabilidad es

$$p(x) = \frac{1}{8\sqrt{x}}, \quad 0 < x < 16 \\ = 0 \text{ en caso contrario}$$

y la función de distribución

$$F(x) = 0, \quad x \leq 0 \\ = \int_0^x \frac{dx}{8\sqrt{x}} = \frac{\sqrt{x}}{4}, \quad 0 < x < 16 \\ = 1, \quad x \geq 16$$

curvas representaciones gráficas son



Finalmente,

$$P[X \geq 8] = p\{X \in [8, 16]\} = \int_8^{16} p(x)dx = \\ = \frac{1}{8} \int_8^{16} \frac{dx}{\sqrt{x}} = \frac{\sqrt{x}}{4} \Big|_8^{16} = 1 - \frac{\sqrt{8}}{4} = 0,2929$$

18 años pese menos de la media, pero más de los 54,39 Kg de peso medio de una mujer de su misma edad.

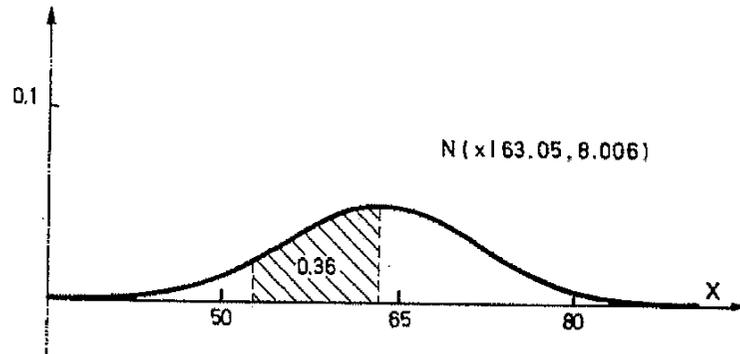
Debido a la simetría, sabemos que el peso medio es

$$\mu = (57,65 + 68,45)/2 = 63,05$$

Además, $p\{x \in [57,65, 68,45]\} = 0,5$, y por simetría $p\{x > 68,45\} = 0,25$. En consecuencia, utilizando (1),

$$\begin{aligned} p\{x > 68,45\} &= p\{x \in [68,45, +\infty]\} = \Phi(\infty) - \Phi\left(\frac{68,45 - 63,05}{\sigma}\right) = \\ &= 1 - \Phi\left(\frac{5,4}{\sigma}\right) = 0,25 \end{aligned}$$

y por tanto $\Phi\left(\frac{5,4}{\sigma}\right) = 0,75$. Buscando en las tablas de la función de distribución normal, o utilizando una calculadora que disponga de esta función, encontramos $\Phi(0,6745) = 0,75$ y, por lo tanto, $5,4/\sigma = 0,6745$ y $\sigma = 8,006$. En consecuencia, la distribución, en Kg, del peso de los varones de 18 años es $N(x|63,05, 8,006)$.



Finalmente, utilizando (1), (3) y las tablas de la función de distribución normal, $p\{54,39 < x < 63,05\}$ es igual a

$$\begin{aligned} \Phi\left(\frac{63,05 - 63,05}{8,006}\right) - \Phi\left(\frac{54,39 - 63,05}{8,006}\right) &= \\ = \Phi(0) - \Phi(-1,08) &= 0,5 - [1 - \Phi(1,08)] = \\ = \Phi(1,08) - 0,5 &= 0,8599 - 0,5 = 0,3599 \end{aligned}$$

La distribución normal aparece también como límite de una familia más amplia de distribuciones, las distribuciones *Student*.

DEFINICIÓN 4.3.6. Una cantidad aleatoria continua X tiene una distribución Student si su función de densidad de probabilidad, que denotaremos $S(x|\mu, \sigma, \alpha)$ es, para todo x real de la forma

$$p(x) = S(x|\mu, \sigma, \alpha) = C \left\{ 1 + \frac{1}{\alpha} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}^{-\frac{\alpha+1}{2}}$$

donde $\mu \in R$, $\sigma > 0$ y $\alpha > 0$. El valor de la constante de proporcionalidad es

$$C = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right) \Gamma\left(\frac{1}{2}\right)} \frac{1}{\sigma \sqrt{\alpha}}$$

Puede comprobarse que la distribución normal es un límite de distribuciones Student. Específicamente,

$$\lim_{\alpha \rightarrow \infty} S(x|\mu, \sigma, \alpha) = N(x|\mu, \sigma)$$

4.4. Funciones de una cantidad aleatoria

Toda función de una cantidad aleatoria (*) es una cantidad aleatoria, puesto que también asocia un número real a cada elemento del conjunto de referencia.

Si X es una cantidad aleatoria discreta e $Y = f(X)$ una función suya, la función de probabilidad de Y se obtiene fácilmente a partir de la de X . En efecto,

$$p(y) = p[Y = y] = p[f(X) = y] = \sum_A p(x) \quad (1)$$

donde se suma para todos los elementos x pertenecientes al conjunto $A = \{x; f(x) = y\}$.

Ejemplo 4.4.1. Pacientes hospitalizados

La distribución de la cantidad aleatoria X que describe el número de pacientes que estarán hospitalizados mañana en una determinada sala es de la forma,

(*) Realmente, toda función medible como, por ejemplo, una biyección.

$$\begin{aligned} p(X = 0) &= 0,1 \\ p(X = 1) &= 0,2 \\ p(X = 2) &= 0,2 \\ p(X = 3) &= 0,4 \\ p(X \geq 4) &= 0,1 \end{aligned}$$

Determinar la distribución de la cantidad aleatoria que describe si hay más de una cama ocupada.

Una función $y = f(x)$ que describe si hay o no más de una cama ocupada es la definida por

$$f(0) = f(1) = 0, \quad f(x) = 1 \quad \text{para todo } x \geq 2$$

y su distribución de probabilidad es

$$\begin{aligned} p(y = 0) &= p(x = 0) + p(x = 1) = 0,3 \\ p(y = 1) &= p(x = 2) + p(x = 3) + p(x \geq 4) = 0,7 \end{aligned}$$

Si X es una cantidad aleatoria continua el problema es más complicado.

TEOREMA 4.4.1. Sea X una cantidad aleatoria continua, sea $p_X(x)$ su función de densidad de probabilidad y sea $Y = f(X)$ una función cuya continua y estrictamente creciente o estrictamente decreciente. Entonces, la función de densidad de Y viene dada por la relación

$$p_Y(y) = p_X[g(y)] \left| \frac{dg(y)}{dy} \right| = p_X(x) \left| \frac{df(x)}{dx} \right| \Big|_{x=g(y)}$$

donde $g(y)$ es la función inversa de $f(x)$, esto es tal que si $y = f(x)$ entonces $x = g(y)$.

Demostración

Supongamos que f es creciente, entonces,

$$F_Y(y) = p[Y \leq y] = p[f(X) \leq y] = p[X \leq g(y)] = F_X[g(y)]$$

derivando con respecto a y resulta, en virtud del Teorema 4.3.1 (ii),

$$p_Y(y) = p_X[g(y)] \frac{dg(y)}{dy}$$

como queríamos demostrar. Si f es decreciente, tenemos

$$F_Y(y) = p[f(X) \leq y] = p[X \geq g(y)] = 1 - F_X[g(y)]$$

y por tanto

$$p_Y(y) = -p_X[g(y)] \frac{dg(y)}{dy}$$

lo que completa la demostración.

Si una función es biyectiva pero no es estrictamente creciente o decreciente, puede descomponerse en secciones en que lo sea y aplicar el Teorema 4.4.1 a cada una de ellas.

Ejemplo 4.4.2. Normalización de una distribución Beta

Sea X una cantidad aleatoria con una distribución Beta, $Be(x|\alpha, \beta)$. Determinar la función de densidad de la cantidad aleatoria $Y = \log\{X/(1-X)\}$ (*).

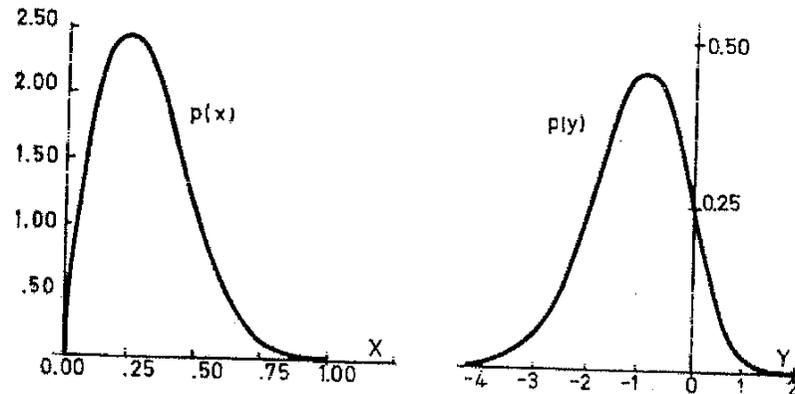
Es fácil comprobar que $y = \log\{x/(1-x)\}$ es una función estrictamente creciente (su derivada es positiva). Además,

$$y = \log \frac{x}{1-x} \rightarrow e^y = \frac{x}{1-x} \rightarrow x = \frac{e^y}{1+e^y}$$

de forma que su función inversa es $x = e^y/(1+e^y)$, cuya derivada es $dx/dy = e^y/(1+e^y)^2$. En consecuencia, como la función de densidad de x es $p(x) = Cx^{\alpha-1}(1-x)^{\beta-1}$ donde $C = \Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)$, la función de densidad de Y será, en virtud del Teorema 4.4.1,

$$\begin{aligned} p(y) &= C \left(\frac{e^y}{1+e^y} \right)^{\alpha-1} \left(\frac{1}{1+e^y} \right)^{\beta-1} \frac{e^y}{(1+e^y)^2} \\ &= C \frac{(e^y)^\alpha}{(1+e^y)^{\alpha+\beta}}, \quad -\infty < y < \infty \end{aligned}$$

La cantidad aleatoria X está definida en $(0, 1)$ mientras que Y lo está de \mathbb{R} . En la figura se hallan representados ambas densidades para $\alpha = 2$ y $\beta = .5$, en cuyo caso, $C = 30$.



(*) Aquí, como en todo el libro, se emplean logaritmos neperianos, esto es de base e , de forma que, para cualquier número real x , $x = \exp\{\log(x)\}$. Se sabe que $e \approx 2,7183$.

Puede observarse que la densidad de probabilidad de Y se parece más a una normal que la de X . Este hecho se utiliza, como veremos, para obtener aproximaciones a probabilidades asociadas a una distribución Beta, utilizando las tablas de la función de distribución normal.

Supongamos que X es una cantidad aleatoria continua, cuya función de distribución es F_X . Entonces, a la cantidad aleatoria $Y = F_X(X)$ se llama su *transformada integral de probabilidad*.

TEOREMA 4.4.2. *La transformada integral de probabilidad $Y = F_X(X)$ de una cantidad aleatoria continua X es una cantidad aleatoria continua con distribución uniforme en $[0, 1]$.*

Demostración

En efecto, $y = F_X(x)$ es una función estrictamente creciente de x , cuya derivada es $p_X(x)$, la función de densidad de x . En consecuencia, utilizando el Teorema 4.4.1 y teniendo en cuenta que $df^{-1}(y)/dy = (df(x)/dx)^{-1}$

$$p_Y(y) = p_X(x)/p_X(x) = 1, \text{ si } 0 \leq y \leq 1$$

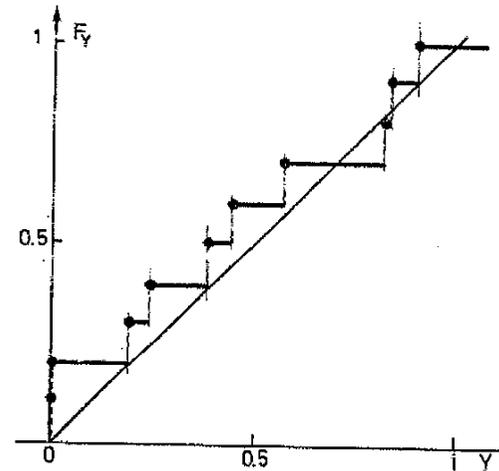
esto es, una densidad uniforme sobre $[0, 1]$.

El Teorema 4.4.2 puede utilizarse para comprobar una hipótesis sobre la distribución de una cantidad aleatoria. En efecto, si $\{x_1, x_2, \dots, x_n\}$ fuesen observaciones de una población con función de distribución F_X , los valores $\{y_1, y_2, \dots, y_n\}$ con $y_i = F_X(x_i)$ serían observaciones de una distribución uniforme en $[0, 1]$. Por otra parte, la representación gráfica de la función de distribución de una cantidad aleatoria Y uniforme en $[\alpha, \beta]$ es, según hemos visto, la recta $F(y) = (y - \alpha)/(\beta - \alpha)$ y por tanto, en nuestro caso, $F(y) = y$; en consecuencia, una forma de comprobar que los valores $\{x_1, \dots, x_n\}$ provienen de una población con función de distribución F_X es representar los valores de la función de distribución correspondiente a los y_i 's y comprobar que, aproximadamente, se sitúan sobre la recta $F(y) = y$.

Ejemplo 4.4.3. Test de normalidad

Comprobar si los valores $-0,73, -0,87, 1,18, -2,09, -0,32, 0,90, -0,16, 0,15, -1,87$ y $0,87$ provienen de una distribución normal $N(x|0, 1)$.

Los correspondientes valores $y_i = \Phi(x_i)$ resultan ser $0,23, 0,19, 0,88, 0,018, 0,37, 0,82, 0,44, 0,56, 0,031$ y $0,81$; asociando la misma probabilidad, esto es $1/10$, a cada uno de estos 10 valores resulta una cantidad aleatoria discreta cuya función de distribución, a la que se llama función de distribución *empírica* del conjunto $\{y_1, y_2, \dots, y_n\}$, resultaría ser



Puede comprobarse que los diez puntos $\{y_i, P(Y \leq y_i)\}$ determinados por la función de distribución empírica de los y_i se sitúan aproximadamente alrededor de la recta $F(y) = y$. El ajuste es razonablemente bueno teniendo en cuenta que sólo se han utilizado diez datos (*) por lo que no existe motivo para rechazar la hipótesis de normalidad.

Supongamos ahora que X es una cantidad aleatoria continua cuya función de distribución es F_X , que F_Y es otra función de distribución, y que pretendemos encontrar una función $f(X)$ cuya función de distribución sea precisamente F_Y . Esto nos permitiría *simular* observaciones de cualquier distribución a partir de observaciones de una distribución conocida.

TEOREMA 4.4.3. *Si X es una cantidad aleatoria continua cuya función de distribución es F_X , la función de distribución de $F_Y^{-1}[F_X(X)]$ es precisamente F_Y .*

Demostración

Sea $z = F_Y^{-1}[F_X(x)]$; entonces

$$F_Z(z) = P\{Z \leq z\} = P\{F_Y^{-1}[F_X(X)] \leq z\} = P\{F_X(X) \leq F_Y(z)\} = F_Y(z)$$

puesto que, en virtud del Teorema 4.4.2, $W = F_X(X)$ tiene una distribución uniforme en $[0, 1]$ cuya función de distribución (Ecuación 4.3.1) es $F_W(x) = x$. En consecuencia, como queríamos demostrar $F_Z = F_Y$.

(*) El análisis cuantitativo de la bondad de este ajustado es un tema importante, pero rebasa los límites impuestos a este volumen. Un análisis Bayesiano de este problema, es el de Bernardo (1980 b).

El resultado que acabamos de demostrar permite *simular* observaciones de una distribución cualquiera utilizando *números aleatorios*, esto es series de números en los que cada uno de los diez dígitos 0, 1, ..., 9 aparece con probabilidad 1/10; este método de simulación es generalmente conocido como *método de Montecarlo*. En efecto, agrupando dígitos aleatorios pueden simularse observaciones x_i de una distribución uniforme en [0, 1], que mediante la función $Y = F^{-1}(X)$ darán lugar, en virtud del Teorema 4.4.3 a observaciones y_i que simulan las que se obtendrían a partir de una distribución cuya función de distribución fuese precisamente F . Se han publicado numerosas tablas de números aleatorios.

Ejemplo 4.4.4. Simulación de observaciones normales

Obtener por simulación cuatro observaciones independientes de una distribución $N(x|0, 1)$.

Agrupados de cuatro en cuatro los dígitos aleatorios de una determinada tabla, encontramos

2369, 8971; 2314, 4806

Considerados como observaciones en [0, 1] tenemos pues los números 0,2369, 0,8971, 0,2314, 0,4806. Sus imágenes inversas en la Tabla de la función de distribución normal son aproximadamente, teniendo en cuenta la relación $\Phi(-x) = 1 - \Phi(x)$,

-0,716, 1,265, -0,734, -0,0486

Estos números por lo tanto pueden utilizarse como si fueran 4 observaciones independientes de una distribución $N(x|0, 1)$.

4.5. Características de una distribución

Lo más relevante de la información contenida en la función de distribución de una cantidad aleatoria puede sintetizarse mediante unos números, sus *características*, cuya observación nos dé una idea aproximada de la configuración de la cantidad aleatoria cuya distribución estudiamos. Es cierto que sólo retenemos con ellos una parte de la información contenida en la función de distribución, pero se trata de la parte más útil.

DEFINICIÓN 5.5.1. Sea X una cantidad aleatoria y sea $f(X)$ una función suya. La media o esperanza de $f(X)$, que denotaremos $E[f(X)]$ es, si existe el valor de la expresión

$$E[f(X)] = \sum f(X_i) p(X_i)$$

si X es una cantidad aleatoria discreta y el de

$$E[f(X)] = \int_{-\infty}^{\infty} f(x) p(x) dx$$

si es continua.

En realidad una definición formal del valor esperado de una cantidad aleatoria exige que la serie o la integral que la definen sean *absolutamente convergentes*, esto es, que $\sum |f(x_i)| p(x_i) < \infty$ o $\int |f(x)| p(x) dx < \infty$. Todos los ejemplos utilizados en este libro cumplen esa condición.

Obsérvese la analogía entre ambas fórmulas. En el caso discreto se trata de una suma sobre todos los valores posibles de $f(X)$ multiplicados por sus probabilidades; en el caso continuo el sumatorio se transforma en integral y la función de probabilidad en función de densidad de probabilidad. De hecho, es posible dar una definición unificada utilizando conceptos de *teoría de la medida*.

El concepto de valor esperado es el que nos permite definir *características* de la distribución de una cantidad aleatoria que constituyen una descripción aproximada de su configuración.

DEFINICIÓN 4.5.2. Los momentos absolutos de orden k , m_k , y los momentos centrales de orden k , μ_k , de una cantidad aleatoria X son los números

$$m_k = E[X^k], \quad \mu_k = E[\{X - E(X)\}^k]$$

En particular, $m_1 = E[X]$ es la media y $\mu_2 = D^2[X]$ la varianza de la distribución de X .

A $D[X]$, la raíz cuadrada de la varianza, se le llama desviación típica y a $D[X]/E[X]$ coeficiente de variación.

La media de una distribución es una medida de su *localización* y puede ser interpretada como el centro de gravedad de la distribución. La *desviación típica* de una distribución es una medida de su *dispersión*, que se mide en las mismas unidades que X y que $E[X]$. El *coeficiente de variación* es una medida adimensional de la dispersión de la distribución.

Tanto la media como la varianza, y por tanto la desviación típica, de una distribución pueden no existir, pero cuando existen su valor es único.

TEOREMA 4.5.1. (Linealidad del operador esperanza). Para toda cantidad aleatoria X y para todo a, b , se verifica que

$$(i) E[ax + b] = aE[X] + b$$

y, como consecuencia,

$$(ii) D^2[aX + b] = a^2 D^2[X]$$

Demostración

Por definición, si X es una cantidad aleatoria discreta,

$$\begin{aligned} E[aX + b] &= \sum (ax_i + b) p(x_i) = a \sum x_i p(x_i) + b \sum p(x_i) = \\ &= a \sum x_i p(x_i) + b = aE[X] + b \end{aligned}$$

la demostración para el caso continuo es análoga.

Además, utilizando este resultado,

$$\begin{aligned} D^2[aX + b] &= E\{(aX + b) - E(aX + b)\}^2 = \\ &= E\{aX + b - aE(X) - b\}^2 = \\ &= E\{a(X - E(X))\}^2 = a^2 E\{X - E(X)\}^2 = a^2 D^2[X] \end{aligned}$$

Los momentos centrales se pueden poner en función de los momentos absolutos; en particular,

TEOREMA 4.5.2. Para toda cantidad aleatoria X ,

$$D^2[X] = E[X^2] - E^2[X]$$

Demostración

Por definición

$$D^2[X] = E\{X - E(X)\}^2 = E\{X^2 - 2E(X)X + E^2(X)\}$$

y, utilizando el Teorema 4.5.1,

$$D^2[X] = E[X^2] - 2E(X)E(X) + E^2(X) = E[X^2] - E^2[X]$$

La media no es la única medida de localización de una distribución de probabilidad ni la varianza, o su raíz cuadrada la desviación típica, es su única medida de dispersión.

DEFINICIÓN 4.5.3. Un cuantil de orden p de una cantidad aleatoria X es una solución de la ecuación $F(x) = p$, donde $F(x)$ es la función de distribución de X . A un cuantil de orden $1/2$ se le llama mediana. Al intervalo definido por los cuantiles de órdenes p y $1 - p$ se le llama intervalo intercuantílico de orden p .

La mediana es una medida de localización de la distribución. Si existe, es el punto que divide los valores posibles de la cantidad aleatoria en dos conjuntos equiprobables; sin embargo, la mediana puede no existir.

Los intervalos intercuantílicos son medidas de dispersión. El intervalo intercuantílico, frecuentemente utilizado, es el intervalo intercuantílico de orden $1/4$.

DEFINICIÓN 4.5.4. Una moda de la distribución de una cantidad aleatoria es un valor que maximiza la función de probabilidad, si X es discreta, o la función de densidad de probabilidad, si X es continua.

La moda de una cantidad aleatoria, que es su valor más probable, es otra medida de localización. Cuando existe, no es necesariamente única.

Ejemplo 4.5.1. Elección de pacientes (cont.)

Determinar las características principales de la cantidad aleatoria discreta definida en el Ejemplo 4.2.1, cuya función de probabilidad es

$$p(X = 0) = 7/15, \quad p(X = 1) = 7/15, \quad p(X = 2) = 1/15$$

Por definición, la media será

$$E[X] = \sum x_i p(x_i) = 0 \cdot \frac{7}{15} + 1 \cdot \frac{7}{15} + 2 \cdot \frac{1}{15} = \frac{9}{15} = 0,6$$

y, análogamente

$$E[X^2] = \sum x_i^2 p(x_i) = 0^2 \cdot \frac{7}{15} + 1^2 \cdot \frac{7}{15} + 2^2 \cdot \frac{1}{15} = \frac{11}{15}$$

En consecuencia, la varianza será

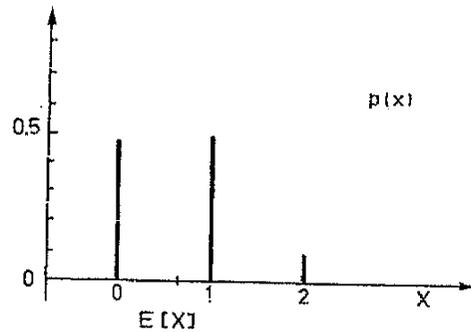
$$D^2[X] = E[X^2] - E^2[X] = \frac{11}{15} - \left(\frac{9}{15}\right)^2 = \frac{156}{225}$$

la desviación típica

$$D[X] = \left(\frac{156}{225}\right)^{1/2} \approx 0,8327$$

y el coeficiente de variación

$$D[X]/E[X] = 0,8327/0,6 \approx 1,3878$$



Existen dos modas, $x = 0$ y $x = 1$ y no existe mediana.

Como puede observarse en el ejemplo anterior, la media de la distribución de una cantidad aleatoria, a diferencia de la moda, no tiene necesariamente que coincidir con alguno de sus valores posibles.

Ejemplo 4.5.2. Tiempos de espera

Se sabe que el tiempo que permanecerá el quirófano de una policlínica sin ser ocupado es una cantidad aleatoria cuya densidad de probabilidad es de la forma

$$p(x) = Ce^{-3x}, \quad \text{si } x > 0$$

$$= 0, \quad \text{para cualquier otro } x.$$

Determinar sus características principales.

En primer lugar debemos determinar la constante C de proporcionalidad. Puesto que $\int p(x)dx = 1$ y

$$\int_0^{\infty} e^{-3x} dx = -\frac{1}{3} e^{-3x} \Big|_0^{\infty} = \frac{1}{3}$$

tenemos $C = \left[\int_0^{\infty} e^{-3x} dx \right]^{-1} = 3$. Además, por definición, la *media* será

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx = 3 \int_0^{\infty} xe^{-3x} dx = \frac{1}{3}$$

y, análogamente,

$$E[X^2] = \int_{-\infty}^{\infty} x^2 p(x) dx = 3 \int_0^{\infty} x^2 e^{-3x} dx = \frac{2}{9}$$

En consecuencia, la *varianza* será

$$D^2[X] = E[X^2] - E^2[X] = \frac{1}{9}$$

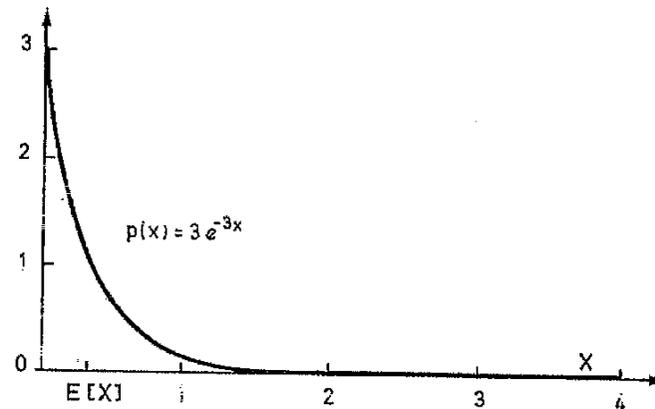
y por tanto, la *desviación típica* $D[X] = 1/3$ y el *coeficiente de variación* $D[X]/E[X] = 1$.

La función $p(x) = 3e^{-3x}$ no tiene ningún máximo y por tanto no existe ninguna moda. La función de distribución es

$$F(x) = 0, \quad x \leq 0$$

$$F(x) = \int_0^x 3e^{-3x} dx = -e^{-3x} \Big|_0^x = 1 - e^{-3x}, \quad x > 0$$

La *mediana*, o cuantil de orden 0,5, es la solución de la ecuación $1 - e^{-3x} = 0,5$, esto es $x = -\log(0,5)/3 = 0,231$



Es conveniente disponer de algunas de las características de las distribuciones más importantes.

TEOREMA 4.5.3 (Características de algunas distribuciones) La definición y los primeros momentos de las distribuciones más comunes son las que se especifican en la Tabla

Distribución	Domnio	$f(x)$	$E[X]$	$D^2[X]$
$Br(x n, \theta)$	$x = 0, 1$	$\theta^x(1-\theta)^{n-x}$	0	$\theta(1-\theta)$
$Bi(x n, \theta, n)$	$x = 0, 1, \dots, n$	$\binom{n}{x} \theta^x(1-\theta)^{n-x}$	$n\theta$	$n\theta(1-\theta)$
$Po(x \lambda)$	$x = 0; 1, 2, \dots$	$e^{-\lambda} \lambda^x / x!$	λ	λ
$Un(x \alpha, \beta)$	$0 < x < \beta$	$1/(\beta - \alpha)$	$(\beta + \alpha)/2$	$(\beta - \alpha)^2/12$
$Be(x \alpha, \beta)$	$0 < x < 1$	$C x^{\alpha-1} (1-x)^{\beta-1}$	$\alpha/(\alpha + \beta)$	$\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$
$Ga(x \mu, \sigma)$	$0 < x < \infty$	$C x^{\alpha-1} e^{-\beta x}$	α/β	α/β^2
$N(x \mu, \sigma)$	$-\infty < x < \infty$	$C \exp\{-\frac{1}{2}[(x-\mu)^2/\sigma^2]\}$	μ	σ^2
$St(x \mu, \sigma, \alpha)$	$-\infty < x < \infty$	$C \left\{ 1 + \frac{1}{\alpha} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}^{-(\alpha+1)/2}$	μ	$\frac{\alpha}{\alpha-2} \sigma^2$

Características de algunas distribuciones

Además, la moda de $Be(\theta|\alpha, \beta)$ es $(\alpha - 1)/(\alpha + \beta - 2)$ y la de $Ga(\theta|\alpha, \beta)$ es $(\alpha - 1)/\beta$. Las modas de $N(x|\mu, \sigma)$ y $St(x|\mu, \sigma, \alpha)$ coinciden con sus medias, por tratarse de distribuciones simétricas.

La demostración puede encontrarse, por ejemplo, en Raiffa & Schlaifer (1961, 3.ª parte).

En la Sección 4.4 estudiamos la forma de determinar la distribución de una función $Y = f(X)$ de una cantidad aleatoria X . Sin embargo, si lo que necesitamos es tan sólo una idea aproximada de sus características, no es necesario determinar toda la distribución para calcularlas. En efecto,

TEOREMA 4.5.4. Sea X una cantidad aleatoria y sea $Y = f(X)$ una función cuya suficientemente regular. Entonces,

- (i) $E[Y] = f\{E[X]\} + \frac{1}{2} D^2[X] f''\{E[X]\}$
- (ii) $D^2[Y] = [f'\{E[X]\}]^2 D^2[X]$

Demostración

Desarrollando $f(X)$ en serie de Taylor alrededor del punto $E[X]$,

$$f(X) = f\{E[X]\} + (X - E[X]) f'\{E[X]\} + \frac{1}{2} (X - E[X])^2 f''\{E[X]\} + \dots$$

y puesto que la esperanza es un operador lineal, tomando valores esperados en ambos lados obtenemos (i).

Además, la ecuación anterior puede reescribirse en la forma

$$Y - f\{E[X]\} = \frac{1}{2} D^2[X] f''\{E[X]\} + (X - E[X]) f'\{E[X]\} + \frac{1}{2} [(X - E[X])^2 - D^2[X]] f''\{E[X]\} + \dots$$

y, en virtud de (i), el primer miembro de la ecuación es aproximadamente $Y - E[Y]$. Elevando al cuadrado, quedándose solo con los términos de orden $(X - E[X])^2$, y tomando esperanzas, se obtiene (ii).

Ejemplo 4.5.3. Normalización de una distribución Beta (cont.)

Sea X una cantidad aleatoria con una distribución Beta, $Be(x|\alpha, \beta)$ y considérese la función $Y = \log\{X/(1 - X)\}$. De acuerdo con lo expuesto en el

Ejemplo 4.4.2 la distribución de Y será, aproximadamente, $N(y|E[Y], D[Y])$. Determinar los valores aproximados de $E[Y]$ y de $D^2[Y]$.

Sabemos (Teorema 4.5.3) que $E[X] = \alpha/(\alpha + \beta)$ y que $D^2[X] = \alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$. Además,

$$y = \log \frac{x}{1-x} \rightarrow y' = \frac{1}{x(1-x)} \rightarrow y'' = \frac{2x-1}{x^2(1-x)^2}$$

En consecuencia, utilizando el Teorema 4.5.5

$$E[Y] = \log \frac{E[X]}{1-E[X]} + \frac{1}{2} D^2[X] \frac{2E[X]-1}{E[X]^2(1-E[X])^2}$$

$$D^2[Y] \approx \left(\frac{1}{E[X](1-E[X])} \right)^2 D^2[X]$$

y, sustituyendo

$$E[Y] \approx \log(\alpha/\beta) + (\alpha - \beta)/2\alpha\beta$$

$$D^2[Y] \approx (\alpha + \beta)/\alpha\beta$$

En general, el cálculo directo de los momentos de una distribución a partir de su definición es complicado. Existe un método más eficaz, basado en la *función generatriz* de momentos.

DEFINICIÓN 4.5.5. La función generatriz de momentos de una cantidad aleatoria X , que representaremos con la letra ψ , es $\psi(t) = E[e^{tX}]$.

La función generatriz no existe necesariamente para todos los valores de t , aunque siempre existe para $t = 0$, puesto que entonces $\psi(0) = E[e^{0X}] = E[1] = 1$. Cuando la función generatriz existe para todos los valores de t en un intervalo que contenga al punto $t = 0$, puede utilizarse para determinar los momentos absolutos de la distribución (*).

TEOREMA 4.5.5. Para toda cantidad aleatoria cuyo momento absoluto de orden k exista,

$$m_k = E[X^k] = \psi^{(k)}(0)$$

donde $\psi^{(k)}(0)$ es el valor de la k -ésima derivada de la función generatriz en el punto $t = 0$.

(*) La función característica, $E[e^{-itX}]$, donde i es la unidad imaginaria, existe siempre y es una generalización de la función generatriz de momentos. Su estudio, sin embargo, sobrepasa los límites impuestos a este libro.

Demostración

En virtud de la definición de la función e^x ,

$$\Psi(t) = E[e^{tX}] = E \left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!} \right] =$$

puesto que la esperanza es un operador lineal,

$$= \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k] = \sum_{k=0}^{\infty} \frac{t^k}{k!} m_k = 1 + tm_1 + \frac{t^2}{2!} m_2 + \frac{t^3}{3!} m_3 + \dots$$

Derivando,

$$\Psi'(t) = m_1 + tm_2 + \frac{t^2}{2!} m_3 + \frac{t^3}{3!} m_4 + \dots$$

$$\Psi''(t) = m_2 + 2tm_3 + \frac{t^2}{2!} m_4 + \frac{t^3}{3!} m_5 + \dots$$

$$\Psi^{(k)}(t) = m_k + tm_{k+1} + \frac{t^2}{2!} m_{k+2} + \frac{t^3}{3!} m_{k+3} + \dots$$

y, por tanto $\Psi'(0) = m_1, \Psi''(0) = m_2, \dots, \Psi^{(k)}(0) = m_k$.

Ejemplo 4.5.4. *Tiempos de espera (cont.)*

Sea X la cantidad aleatoria definida en el Ejemplo 4.5.2, cuya función de densidad de probabilidad era

$$p(x) = 3e^{-3x}, \text{ si } x > 0$$

$$= 0, \text{ para cualquier otro } x.$$

Determinar su función generatriz y utilizarla para determinar $E[X]$ y $D^2[X]$.

Por definición,

$$\Psi(t) = E[e^{tX}] = \int_0^{\infty} e^{tx} 3e^{-3x} dx = \frac{3}{3-t} \quad t < 3$$

la función generatriz no existe, en este caso, para $t \geq 3$.

$$\Psi'(t) = \frac{3}{(3-t)^2}, \quad \Psi'(0) = E[X] = \frac{1}{3}$$

$$\Psi''(t) = \frac{6}{(3-t)^3}, \quad \Psi''(0) = E[X^2] = \frac{2}{9}$$

y por lo tanto $D^2[X] = E[X^2] - E^2[X] = 1/9$.

Bajo condiciones muy generales, la función generatriz caracteriza a la distribución, de forma parecida a como lo hacen la función de distribución o la de (densidad de) probabilidad. El estudio detallado de este tema excede, sin embargo, los límites de este volumen.

4.6. Distribuciones multivariantes

En numerosas situaciones es necesario considerar simultáneamente las propiedades de dos o más cantidades aleatorias. Por ejemplo, las consecuencias de un determinado tratamiento pueden depender *simultáneamente* de la presión sanguínea y de la cantidad de azúcar en la sangre; los posibles efectos de una determinada enfermedad hereditaria pueden depender *simultáneamente* del sexo, la edad y el grupo sanguíneo del paciente.

DEFINICIÓN 4.6.1. Dado un espacio probabilístico (Ω, Σ, P) , un vector aleatorio de dimensión k es una función que asocia un elemento de R^k a cada elemento de Ω .

Naturalmente, cada una de los componentes de un vector aleatorio es una cantidad aleatoria. Así, si a un nuevo paciente le asignamos su temperatura, presión sanguínea y edad, hemos definido un vector aleatorio de dimensión tres, cuyos componentes son las cantidades aleatorias mencionadas.

Todas las probabilidades relativas a los valores que puede tomar un vector aleatorio pueden obtenerse directamente a partir de su función de distribución.

DEFINICIÓN 4.6.2. La función de distribución de un vector aleatorio $\mathbf{X} = \{X_1, X_2, \dots, X_k\}$ es la función de R^k en $[0, 1]$ definida mediante

$$F(x_1, x_2, \dots, x_k) = P[\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\} \cap \dots \cap \{X_k \leq x_k\}]$$

Si un vector aleatorio \mathbf{X} puede tomar tan solo un número finito o numerable de valores distintos, decimos que es discreto. Sus componentes son

entonces cantidades aleatorias discretas y su función de probabilidad es la función de R^k en $[0, 1]$

$$p(x_1, x_2, \dots, x_k) = P[\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \dots \cap \{X_k = x_k\}] \quad (1)$$

Por otra parte, si la probabilidad de que un vector aleatorio \mathbf{X} de dimensión k pertenezca a una determinada región A de R^k puede expresarse mediante una integral de la forma

$$P[\mathbf{X} \in A] = \int \int \dots \int_A p(x_1, x_2, \dots, x_k) dx_1 dx_2 \dots dx_k \quad (2)$$

decimos que \mathbf{X} es un vector aleatorio continuo y que $p(x_1, x_2, \dots, x_k)$ es su función de densidad de probabilidad. Sus componentes son entonces cantidades aleatorias continuas, y se verifica que

$$p(x_1, x_2, \dots, x_k) = \frac{\partial^k F(x_1, x_2, \dots, x_k)}{\partial x_1 \partial x_2 \dots \partial x_k} \quad (3)$$

Es obvio que los conceptos relativos a vectores aleatorios son una generalización natural de los comentados en las secciones anteriores para cantidades aleatorias.

Las distribuciones de subconjuntos cualesquiera de las componentes de un vector aleatorio pueden ser determinadas a partir de la distribución del vector y reciben el nombre de distribuciones marginales.

DEFINICIÓN 4.6.3. (Distribuciones marginales). Sea \mathbf{X} un vector aleatorio de dimensión k , sea $(\mathbf{X}_1, \mathbf{X}_2)$ una partición de las componentes de \mathbf{X} y sea l la dimensión de \mathbf{X}_1 .

Si \mathbf{X} es discreto, y su función de probabilidad es $p(x_1, x_2)$, la función de probabilidad de \mathbf{X}_1 es

$$p(\mathbf{X}_1 = x_1) = \sum_{x_2} p(x_1, x_2)$$

donde la suma se efectúa sobre todos los puntos (x_1, x_2) para los que $\mathbf{X}_1 = x_1$.

Si \mathbf{X} es continuo, y su función de densidad de probabilidad es $p(x_1, x_2)$, la función de densidad de probabilidad de \mathbf{X}_1 es

$$p(x_1) = \int_{R^{k-1}} p(x_1, x_2) dx_2$$

Ejemplo 4.6.1. Tipos de enfermedad

Se sabe que un determinado síndrome puede ser causado por tres enfermedades distintas, que cada una de ellas tiene una forma benigna y una maligna, y que el porcentaje de veces en que se presenta cada una de las combinaciones posibles es

	ω_1	ω_2	ω_3
B	30	40	10
M	10	9	1

Determinar la función de probabilidad de un vector aleatorio que describe la situación y las de sus dos distribuciones marginales.

La situación puede ser descrita, por ejemplo, mediante el vector aleatorio $Z = (X, Y)$ donde

$$\begin{aligned} X(B) &= 0, & X(M) &= 1 \\ Y(\omega_i) &= i, & i &= 1, 2, 3 \end{aligned}$$

la correspondiente función de probabilidad será claramente

$$\begin{aligned} p(0, 1) &= 0,30, & p(0, 2) &= 0,40, & p(0, 3) &= 0,10 \\ p(1, 1) &= 0,10, & p(1, 2) &= 0,09, & p(1, 3) &= 0,01 \end{aligned}$$

la marginal de la cantidad aleatoria X que describe si la enfermedad se presenta en forma benigna o maligna es

$$\begin{aligned} P(x=0) &= p(0, 1) + p(0, 2) + p(0, 3) = 0,80 \\ P(x=1) &= p(1, 1) + p(1, 2) + p(1, 3) = 0,20 \end{aligned}$$

de forma que el 80% de los casos son benignos.

La marginal de la cantidad aleatoria Y que describe la enfermedad es

$$\begin{aligned} P(Y=1) &= p(0, 1) + p(1, 1) = 0,40 \\ P(Y=2) &= p(0, 2) + p(1, 2) = 0,49 \\ P(Y=3) &= p(0, 3) + p(1, 3) = 0,11 \end{aligned}$$

de forma que la enfermedad más frecuente es ω_2 que se presenta en el 49% de los casos.

En general, el valor de unas componentes de un vector aleatorio influye sobre el valor de otras. Esto no sucede cuando son *independientes*.

DEFINICIÓN 4.6.4. Se dice que las componentes de un vector aleatorio \mathbf{X} son independientes entre sí, si se verifica que

$$p(x_1, \dots, x_k) = p(x_1) p(x_2) \dots p(x_k)$$

es decir su función de probabilidad, o de densidad de probabilidad, es igual al igual al producto de las correspondientes funciones marginales.

La independencia es un caso particular de un concepto más general y mucho más frecuente en la práctica, la *intercambiabilidad*.

DEFINICIÓN 4.6.5. Se dice que las componentes de un vector aleatorio \mathbf{X} son intercambiables entre sí, si para toda posible permutación suya $(X_{i_1}, X_{i_2}, \dots, X_{i_k})$ se verifica que

$$p(x_1, \dots, x_k) = p(x_{i_1}, \dots, x_{i_k})$$

de forma que su función de probabilidad o de densidad de probabilidad es independiente del orden en que se sitúan.

Intuitivamente, un conjunto de cantidades aleatorias son intercambiables cuando las conclusiones que pueden extraerse de su observación no dependen del orden en que han sido observadas.

La influencia de unas cantidades aleatorias sobre otras puede ser determinada si se conoce la distribución del vector aleatorio que se obtiene tomándolas como componentes.

DEFINICIÓN 4.6.6. (Distribuciones condicionales.) Sea \mathbf{X} un vector aleatorio de dimensión k , sea $(\mathbf{X}_1, \mathbf{X}_2)$ una partición de las componentes de \mathbf{X} y sea i la dimensión de \mathbf{X}_1 . Si $p(x_1, x_2)$ es la función de probabilidad, en el caso discreto, o la función de densidad de probabilidad, en el caso continuo, del vector \mathbf{X} , la distribución de \mathbf{X}_1 cuando se sabe que $\mathbf{X}_2 = x_2$ viene dada por

$$p(x_1|x_2) = p(x_1, x_2)/p(x_2)$$

La definición de las distribuciones condicionales está claramente motivada por el Teorema de Bayes. Naturalmente, si \mathbf{X}_1 y \mathbf{X}_2 son independientes, y solo entonces, resulta que $p(x_1|x_2) = p(x_1)$.

Ejemplo 4.6.2. Tipos de enfermedad (cont.)

Determinar las distribuciones condicionales correspondientes al vector aleatorio del Ejemplo 4.6.1.

Las distribuciones condicionales de la cantidad aleatoria X , que determina si la enfermedad se presenta en forma benigna, dada la enfermedad, serán

$$\left\{ \begin{aligned} p(x=0|y=1) &= \frac{p(0, 1)}{p(y=1)} = \frac{0,30}{0,40} = 0,750 \\ p(x=1|y=1) &= 1 - p(x=0|y=1) = 0,250 \\ p(x=0|y=2) &= \frac{p(0, 2)}{p(y=2)} = \frac{0,40}{0,49} = 0,816 \\ p(x=1|y=2) &= 1 - p(x=0|y=2) = 0,184 \\ p(x=0|y=3) &= \frac{p(0, 3)}{p(y=3)} = \frac{0,10}{0,11} = 0,909 \\ p(x=1|y=3) &= 1 - p(x=0|y=3) = 0,091 \end{aligned} \right.$$

Análogamente, las distribuciones condicionales de la cantidad aleatoria Y , que determina la enfermedad, sabiendo si la forma es benigna o maligna serán

$$\left\{ \begin{aligned} p(y=1|x=0) &= \frac{p(0, 1)}{p(x=0)} = \frac{0,30}{0,80} = 0,375 \\ p(y=2|x=0) &= \frac{p(0, 2)}{p(x=0)} = \frac{0,40}{0,80} = 0,500 \\ p(y=3|x=0) &= 1 - p(y=1|x=0) - p(y=2|x=0) = 0,125 \\ p(y=1|x=1) &= \frac{p(1, 1)}{p(x=1)} = \frac{0,10}{0,20} = 0,500 \\ p(y=2|x=1) &= \frac{p(1, 2)}{p(x=1)} = \frac{0,09}{0,20} = 0,450 \\ p(y=3|x=1) &= 1 - p(y=1|x=1) - p(y=2|x=1) = 0,050 \end{aligned} \right.$$

Una función de un vector aleatorio es otro vector aleatorio (o una cantidad aleatoria, que no es más que un vector aleatorio de dimensión uno).

La distribución de una función biunívoca de un vector aleatorio puede ser determinada mediante un teorema que generaliza el 4.4.1.

TEOREMA 4.6.1 Sea $\mathbf{X} = \{X_1, \dots, X_k\}$ es un vector aleatorio cuya densidad de probabilidad es $p_X(\mathbf{x})$ y sea $\mathbf{Y} = f(\mathbf{X}) = \{f_1(\mathbf{X}), \dots, f_k(\mathbf{X})\}$ una transformación biunívoca suya. Entonces, la densidad de probabilidad de \mathbf{X} es

$$p_X(\mathbf{y}) = p_X[f^{-1}(\mathbf{y})] |J(\mathbf{y})|$$

donde $\mathbf{x} = g(\mathbf{y}) = \{g_1(\mathbf{y}), \dots, g_k(\mathbf{y})\}$ es la función inversa de $f(\mathbf{x})$, siempre que el Jacobiano

$$J(\mathbf{y}) = \begin{vmatrix} \frac{\partial g_1(\mathbf{y})}{\partial y_1} & \dots & \frac{\partial g_1(\mathbf{y})}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_k(\mathbf{y})}{\partial y_1} & \dots & \frac{\partial g_k(\mathbf{y})}{\partial y_k} \end{vmatrix}$$

exista, no sea nulo, y tenga todas sus derivadas parciales continuas.

La demostración es una generalización inmediata de la del Teorema 4.4.1.

Para determinar la distribución de una función $f_1(\mathbf{x})$ de un vector aleatorio k -dimensional cuya dimensión i es menor que k puede procederse de la forma siguiente: (i) Se determina otra función $f_2(\mathbf{x})$ de dimensión $k-i$, de forma que la transformación conjunta $f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$ sea una transformación biunívoca; (ii) se determina entonces la distribución de $f(\mathbf{x})$ mediante el Teorema 4.6.2 y (iii) se procede finalmente a obtener la distribución marginal de $f_1(\mathbf{x})$.

Ejemplo 4.6.3. Distribución del cociente

Sea $\mathbf{X} = (X_1, X_2)$ un vector aleatorio continuo cuya función de densidad de probabilidad es

$$p(x_1, x_2) = 4x_1x_2, \text{ si } 0 < x_1 < 1 \text{ y } 0 < x_2 < 1 \\ = 0, \text{ en cualquier otro caso.}$$

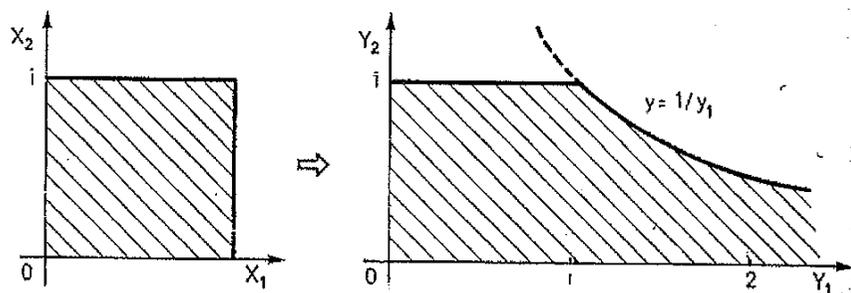
Determinar la distribución de $Y_1 = X_1/X_2$.

De acuerdo con el procedimiento que acabamos de describir, empezamos completando la transformación para hacerla biunívoca. Una sencilla elección es $Y_2 = X_2$.

Si $Y_1 = X_1/X_2$ y $Y_2 = X_2$, entonces $X_1 = Y_1Y_2$, y $X_2 = Y_2$, de forma que la correspondiente función inversa es $g(\mathbf{y}) = \{y_1y_2, y_2\}$ y su Jacobiano resulta ser

$$J(\mathbf{y}) = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2$$

La región $0 < x_1 < 1, 0 < x_2 < 1$ se transforma en la $0 < y_1 < \infty, 0 < y_2 < 1, 0 < y_1y_2 < 1$.



y por lo tanto la función de densidad de probabilidad de y será

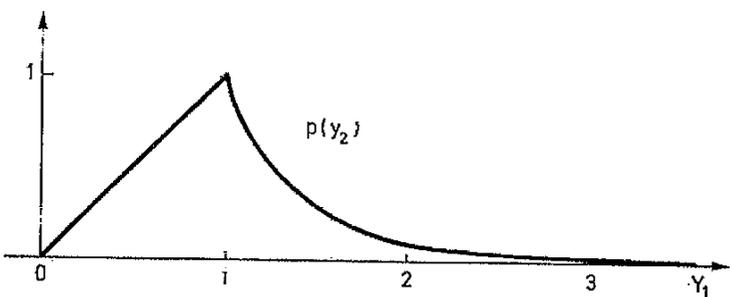
$$p(y_1, y_2) = 4y_1y_2, \quad 0 < y_1 < \infty \quad 0 < y_2 < 1 \quad \vee \quad 0 < y_1y_2 < 1$$

$$= 0, \text{ en cualquier otro caso.}$$

Finalmente, la densidad de probabilidad de y_1 será

$$p(y_1) = \int p(y_1, y_2) dy_2 = 4y_1 \frac{y_2^4}{4} \Big|_0^1 = y_1, \quad 0 < y_1 < 1$$

$$= 4y_1 \frac{y_2^4}{4} \Big|_0^{1/y_1} = \frac{1}{y_1^2}, \quad 1 < y_1 < \infty$$



Los conceptos relativos a las *características* de las cantidades aleatorias se generalizan sin dificultad a los vectores aleatorios. Así, el *valor esperado* de una función $f(X)$ de un vector aleatorio sigue tomando la forma de la Definición 4.5.1, teniendo en cuenta que en el caso continuo se trata ahora de una integral múltiple.

Si $\mathbf{X} = (X_1, X_2, \dots, X_k)$ es un vector aleatorio k -dimensional, su *media* $E[\mathbf{X}]$ es el vector k -dimensional $\{E(X_1), \dots, E(X_k)\}$ cuyas componentes son las medias de las k distribuciones marginales de sus componentes. Su *matriz de varianzas-covarianzas*, $D^2(\mathbf{X})$, que generaliza el concepto de varianza de una cantidad aleatoria, es la matriz simétrica de dimensión k cuyo elemento general es

$$C[X_i, X_j] = E\{[X_i - E[X_i]] [X_j - E[X_j]]\}$$

De esta forma los elementos diagonales de la matriz de varianzas-covarianzas son precisamente las varianzas de las k distribuciones marginales, esto es $C[X_i, X_i] = D^2[X_i]$. Los demás elementos de la matriz reciben el nombre de *covarianzas*. El *coeficiente de correlación* entre las componentes X_i y X_j se define como

$$C[X_i, X_j] / \{D[X_i] D[X_j]\}$$

Interesan frecuentemente las características de la distribución de una combinación lineal de cantidades aleatorias.

TEOREMA 4.6.2. Sea $\mathbf{X} = \{X_1, \dots, X_k\}$ un vector aleatorio cuya esperanza existe. Entonces,

$$E[a_1X_1 + \dots + a_kX_k + b] = a_1E[X_1] + \dots + a_kE[X_k] + b$$

Demostración

$$E[a_1X_1 + \dots + a_kX_k + b] =$$

$$= \iiint [a_1x_1 + \dots + a_kx_k + b] p(x_1, \dots, x_k) dx_1, \dots, dx_k =$$

$$= a_1 \int x_1 p(x_1) dx_1 + \dots + a_k \int x_k p(x_k) dx_k + b =$$

$$= a_1E[X_1] + \dots + a_kE[X_k] + b$$

como queríamos demostrar. La demostración para el caso discreto es análoga.

TEOREMA 4.6.3. Si X_1, \dots, X_k son cantidades aleatorias independientes cuya esperanza existe,

$$E[X_1X_2 \dots X_k] = E[X_1] E[X_2] \dots E[X_k]$$

y para todo par (X_i, X_j) , $C[X_i, X_j] = 0$.

Demostración

Por definición, si las X_i son independientes, $p(x_1, \dots, x_k) = \prod p(x_i)$. En consecuencia, en el caso continuo,

$$E[\prod X_i] = \int \int (\prod x_i) \prod p(x_i) = \prod \int x_i p(x_i) = \prod E[X_i]$$

El mismo resultado se obtiene en el caso discreto. Finalmente,

$$C[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])] = E[X_i - E[X_i]] E[X_j - E[X_j]] = 0$$

TEOREMA 4.6.4. Si X_1, \dots, X_k son cantidades aleatorias independientes cuya varianza existe

$$D^2[a_1 X_1 + \dots + a_k X_k + b] = a_1^2 D^2[X_1] + \dots + a_k^2 D^2[X_k]$$

Demostración

Demostremos primero que $D^2[X_1 + X_2] = D^2[X_1] + D^2[X_2]$. En efecto,

$$\begin{aligned} D^2[X_1 + X_2] &= E\{(X_1 + X_2 - E[X_1] - E[X_2])^2\} \\ &= E\{(X_1 - E[X_1])^2 + (X_2 - E[X_2])^2 + 2(X_1 - E[X_1])(X_2 - E[X_2])\} = \\ &= D^2[X_1] + D^2[X_2] + C[X_1, X_2] = D^2[X_1] + D^2[X_2] \end{aligned}$$

en virtud del Teorema 4.6.4. Extendiendo este resultado a k variables y utilizando el Teorema 4.5.1 (ii) se completa la demostración.

La más utilizada de las distribuciones continuas multivariantes es la *distribución normal multivariante* que generaliza la distribución normal definida en la Sección anterior.

DEFINICIÓN 4.6.7. Un vector aleatorio continuo \mathbf{X} de dimensión k tiene una distribución normal si su función de densidad de probabilidad, que denotaremos $N_k(\mathbf{x}|\mu, \Sigma)$ es de la forma

$$p(x_1, \dots, x_k) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{k/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)' \Sigma^{-1}(\mathbf{x} - \mu)\right\}$$

donde $\mathbf{x}' = \{x_1, \dots, x_k\} \in R^k$, $\mu' = \{\mu_1, \dots, \mu_k\} \in R^k$ y Σ es una matriz definida positiva (*).

(*) Una matriz A es definida positiva si para cualquier vector $\mathbf{x} \neq 0$ se verifica que $\mathbf{x}' A \mathbf{x} > 0$, donde \mathbf{x}' es el vector transpuesto de \mathbf{x} .

Los parámetros μ y Σ contienen, respectivamente, los momentos de primer y segundo orden; en efecto,

TEOREMA 4.6.5. Si \mathbf{X} es un vector aleatorio con distribución normal $N_k(\mathbf{x}|\mu, \Sigma)$, $E[\mathbf{X}] = \mu$ y $D^2[\mathbf{X}] = \Sigma$.

La demostración puede consultarse, por ejemplo en Raiffa & Schlaifer (1961, Cap. 8).

Cuando $k = 2$, la correspondiente densidad normal *bivariante* es una superficie acampanada centrada en el punto $(x_1, x_2) = (\mu_1, \mu_2)$. Para facilitar la interpretación de los resultados, la normal bivariante se suele describir como función de sus medias μ_1, μ_2 , desviaciones típicas σ_1, σ_2 y coeficiente de correlación ρ . De esta forma, con

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (4)$$

la función de densidad de la normal bivariante resulta ser

$$\begin{aligned} N(x_1, x_2|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) &= \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - \right. \right. \\ &\quad \left. \left. - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right] \right\} \end{aligned}$$

Volviendo al caso general,

TEOREMA 4.6.6. Sea \mathbf{X} un vector aleatorio con distribución normal $N_k(\mathbf{x}|\mu, \Sigma)$ y sea (X_1, X_2) una partición de las componentes de \mathbf{X} , sea i la dimensión X_1 ,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

las correspondientes particiones de μ y Σ . Entonces

$$p(x_1) = N_i(x_1|\mu_1, \Sigma_{11})$$

$$p(x_1|x_2) = N_i(x_1|\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

La demostración puede consultarse, por ejemplo, en Raiffa & Schlaifer (1961, Cap. 8). Cuando $k = 2$, resulta

TEOREMA 4.6.7. Sea $\mathbf{X} = (X_1, X_2)$ un vector aleatorio con distribución normal $N(x_1, x_2 | \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Entonces,

$$p(x_1) = N(x_1 | \mu_1, \sigma_1)$$

$$p(x_2) = N(x_2 | \mu_2, \sigma_2)$$

$$p(x_1 | x_2) = N\left(x_1 | \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2), \sigma_1 \sqrt{1 - \rho^2}\right)$$

$$p(x_2 | x_1) = N\left(x_2 | \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1), \sigma_2 \sqrt{1 - \rho^2}\right)$$

Ejemplo 4.6.4. Pesos y estaturas

Se sabe que la distribución conjunta del peso en Kg X_1 , y la estatura en cm X_2 de un varón de 18 años es una distribución normal cuyos parámetros son

$$\mu = \begin{pmatrix} 63,05 \\ 174,5 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 64,09 & 36,55 \\ 36,55 & 42,55 \end{pmatrix}$$

Determinar la probabilidad de que un varón de 18 años y 167 cm de estatura tenga un peso comprendido entre 54,39 y 63,05 Kg. Comparar el resultado con el obtenido en el Ejemplo 4.3.2.

De acuerdo con (4) los parámetros de la distribución conjunta de (X_1, X_2) serán $\mu_1 = 63,05$, $\mu_2 = 174,5$, $\sigma_1 = 8,01$, $\sigma_2 = 6,52$ y $\rho = 0,70$. Consecuentemente, la distribución condicional del peso X_1 , dada la estatura $X_2 = 167$ será, utilizando el Teorema 4.6.7

$$\begin{aligned} N(x_1 | 63,05 + 0,7 \frac{8,01}{6,52} (167 - 174,5), 8,01 \sqrt{1 - 0,7^2}) &= \\ &= N(x_1 | 56,61, 5,718) \end{aligned}$$

Por tanto, utilizando la Ecuación 4.3.1 y las tablas de la Normal,

$$\begin{aligned} p[54,39 < X_1 < 63,05 | X_2 = 167] &= \\ &= \Phi\left(\frac{63,05 - 56,61}{5,718}\right) - \Phi\left(\frac{54,39 - 56,61}{5,718}\right) = \\ &= \Phi(1,126) - \Phi(-0,388) = 0,5211 \end{aligned}$$

Como era de esperar, la probabilidad obtenida, 0,5211 es bastante mayor que la obtenida en el Ejemplo 4.3.2, 0,3599. En efecto, la altura y el peso están positivamente relacionados, de forma que sabiendo que el sujeto tiene una altura de 167 cm, inferior

a la media, es de esperar que su peso sea también inferior a la media, y por tanto que sea más alta la probabilidad de que se sitúe en un intervalo de valores inferiores a la media.

4.7. Discusión y referencias

Los conceptos discutidos en este capítulo son independientes de la interpretación que se dé a la noción de probabilidad. En consecuencia el concepto de cantidad aleatoria, o variable aleatoria como se le llama más frecuentemente, aparece en todos los textos de Teoría de la Probabilidad.

Un tratamiento profundo de las ideas desarrolladas en este capítulo exige ciertos conocimientos de *teoría de la medida*. El uso de tales ideas permite unificar el estudio de las cantidades aleatorias discretas y continuas, pero reduce notablemente el conjunto de personas que pueden seguirlo.

Existen numerosas distribuciones de probabilidad, además de las definidas en las Secciones 4.2 y 4.3, que aparecen frecuentemente en las aplicaciones. Numerosas monografías han sido dedicadas a su estudio sistemático. Entre las más completas pueden citarse los cuatro volúmenes de Johnson & Kotz (1969/1972), que incluyen además una extensa bibliografía, y la tercera parte del texto de Raiffa & Schlaifer (1961).

Entre los textos clásicos que discuten el concepto de cantidad aleatoria merecen mención especial las de Laplace (1812/1912), Hausdorff (1914), Von Mises (1936), Jeffreys (1939/1967), Kolmogorov (1933), Feller (1957/1966), Loeve (1955/1977), Renvi (1962/1966), Gnedenko (1962) y De Finetti (1970/1975) todos ellos de lectura obligada para un estudioso de la Teoría de la Probabilidad.

Casi todos los textos de estadística matemática incluyen una discusión de los conceptos desarrollados en este capítulo. Podemos mencionar entre ellos los de Freeman (1963), Hogg & Craig (1965), Lindgren (1962), Lindley (1965), Mood, Graybill & Boes (1963/1974), Papoulis (1965) y Parzen (1960). A un nivel más avanzado están los textos de Anderson (1958), Ash (1972), Cramer (1946), Feller (1937/1966), Fisz (1963), Kriskberg (1965), Rao (1965), Rohatgi (1976) y Wilks (1962), además de los «clásicos» ya mencionados. El libro de Kingman & Taylor (1966) proporciona una interesante introducción a la teoría de la probabilidad como un caso particular de teoría de la medida.

La existencia de densidades de probabilidad está relacionada con el postulado de σ -aditividad que, como mencionamos en el Capítulo 3, es aceptado por casi todos los autores, con la notable excepción de De Finetti (1970/1975). Este autor es por otra parte responsable del concepto de cantidades aleato-

rias *intercambiables* que, como veremos más adelante, aparece de forma natural en los problemas de inferencia.

Entre los conceptos y resultados que hemos decidido omitir en nuestro breve tratamiento de las cantidades aleatorias, hay que mencionar (i) el concepto de *función característica* que generaliza el de *función generatriz*, (ii) su uso para identificar distribuciones y para determinar la distribución de sumas de cantidades aleatorias independientes y (iii) los problemas de *convergencia* de sucesiones de cantidades aleatorias. Merecen atención especial los *teoremas de límite*, sobre los resultados de tal convergencia, de los que el Teorema 4.2.1 y el *teorema central del límite* mencionado en la Sección 4.3, son los ejemplos más conocidos. Todos estos temas pueden ser consultados en cualquiera de los textos avanzados que hemos citado.

PROBLEMAS

1. El número de operaciones que serán realizadas en el Departamento de Traumatología del Hospital Clínico de la Universidad de Valencia en las próximas 24 horas es una cantidad aleatoria cuya función de probabilidad viene dada por la Tabla

x	p(x)
0	0,10
1	0,15
2	0,35
3	0,20
4	0,15
5	0,05

Representar gráficamente $p(x)$ y su correspondiente función de distribución. Determinar $p[X \geq 2,5]$, $p[0 < X < 2]$, $p[0 < X \leq 2]$, $p[X > 6]$, $p[X = 1,5]$.

2. El número de llamadas que deberán ser atendidas una determinada madrugada por un médico de guardia es una cantidad aleatoria X cuya función de probabilidad es de la forma

$$p(x) = C/(x + 1), \quad x = 0, 1, 2, 3, 4,$$

$$= 0, \quad \text{en otro caso.}$$

Determinar C , dibujar las correspondientes funciones de probabilidad y de distribución y calcular $p[X > 1]$.

3. Las horas transcurridas entre dos urgencias consecutivas es una cantidad aleatoria X cuya función de densidad de probabilidad es de la forma

$$p(x) \propto e^{-ax} \quad x > 0$$

$$= 0, \quad \text{en otro caso.}$$

Calcular la correspondiente constante de proporcionalidad, representar las funciones $p(x)$ y $F(x)$, y determinar $p[X \geq 0,5]$.

4. La cantidad m , en gramos, de proteínas contenidas en 100 ml de plasma sanguíneo de una persona adulta es una cantidad aleatoria X con distribución normal $N(x|6,72, 0,35)$. Calcular $p[X < 6]$ y $p[6 < X < 7]$ y determinar un intervalo centrado en 6,72 cuya probabilidad sea 0,95.
5. Se sabe que la cantidad de potasio contenido en la sangre humana completa es una cantidad aleatoria X con distribución normal centrada en 43 mg/100 ml, y que $p[X \leq 32] = 0,025$. Determinar $D[X]$ y $p[X > 50]$.
6. El logaritmo decimal de las horas necesarias para realizar una determinada operación es una cantidad aleatoria X con distribución normal $N(x|0,5, 0,3)$. Determinar la probabilidad de que la operación dure más de cuatro horas.
7. La probabilidad de sobrevivir una peligrosa operación cerebral es una cantidad aleatoria X con distribución Beta, $Be(x|6, 4)$. Utilizar el hecho de que $Y = \log\{X/(1-X)\}$ tiene entonces una distribución aproximadamente normal para determinar $p[X \geq 0,5]$.
8. En una sala de espera se encuentran 15 pacientes, 5 hombres y 10 mujeres, de los que se hace pasar a los 3 primeros. Sea X el número de mujeres e Y el número de hombres seleccionados. Determinar la distribución y las principales características de la cantidad aleatoria $X - Y$.
9. La anchura de la pevis X_1 y el perímetro torácico X_2 , en cm en mujeres de 18 años tienen una distribución conjunta normal bivalente con

$$E[X] = \begin{pmatrix} 22,0 \\ 63,9 \end{pmatrix} \quad D^2[X] = \begin{pmatrix} 1,09 & 2,28 \\ 2,28 & 9,82 \end{pmatrix}$$

Determinar $p[X_2 > 66]$, $p[20 < X_1 < 22]$ y $p[X_2 > 66|X_1 = 21]$.

10. Las cantidades de ácido cítrico X_1 y de ácido láctico X_2 contenidas en la orina tienen una distribución normal bivalente centrada en 678 y 350 mg/24 h respectivamente, y con un coeficiente de correlación 0,9. Sabiendo que $p[X_1 < 128] = p[X_2 < 100] = 0,02$, determinar $p[X_1 > 800]$ y $p[X_1 > 800|X_2 = 500]$.

El proceso de aprendizaje

En este capítulo se describe el *proceso* que permite incorporar al análisis de un problema de decisión la información proporcionada por datos experimentales relacionados con sus sucesos inciertos relevantes.

Este proceso exige precisar, mediante la *función de verosimilitud*, la relación existente entre los datos y los sucesos inciertos; describir, mediante una *distribución* de probabilidad, la información *inicial* que se posee sobre la verosimilitud de su ocurrencia y determinar, mediante el *Teorema de Bayes*, la *distribución final* que describe la información que se posee sobre los sucesos inciertos tras incorporar a la información inicial la que proporcionan los resultados experimentales.

Este *proceso de aprendizaje* constituye la base de todo problema de *inferencia*: hacer inferencias sobre el valor de θ se reduce básicamente a determinar su distribución final. Los problemas de *predicción* se resuelven como una extensión natural de los problemas de inferencia mediante el uso del teorema de la probabilidad total.

La reacción natural de cualquiera que tenga que tomar una decisión cuyas consecuencias dependen de la magnitud de una cantidad desconocida θ es intentar reducir su incertidumbre obteniendo más información sobre su valor. El problema central de la *inferencia estadística* es el de proporcionar una metodología que permita asimilar la información que resulte accesible, con objeto de mejorar nuestro conocimiento del mundo real.

5.1. Cuantificación de la información inicial

Supongamos, sin pérdida de generalidad, que estamos interesados en el valor de una *cantidad* (*) θ que denominaremos *parámetro de interés*. De acuerdo con los argumentos ofrecidos en el capítulo de Fundamentos, la información disponible sobre el valor de θ deberá ser expresada mediante una medida de probabilidad que describa el grado de creencia del investigador en la ocurrencia de los distintos valores posibles de θ . Se trata pues de una *cantidad aleatoria* cuya distribución de probabilidad describe la información sobre el valor de θ que inicialmente se posee; esta distribución recibe el nombre de *distribución inicial* de θ .

Si θ es una cantidad aleatoria *discreta*, su distribución de probabilidad puede ser descrita mediante la correspondiente función de probabilidad $p(\theta) = \{p_1, p_2, \dots\}$. Las probabilidades $p_i = p(\theta_i)$ pueden ser determinadas mediante el uso de técnicas como las mencionadas en la Sección 3.6, o mediante el establecimiento de relaciones entre ellas que permitan determinarlas.

Ejemplo 5.1.1. *Diagnosis*

Las consecuencias de un determinado tratamiento dependen de la enfermedad del paciente. Se considera que existen cinco enfermedades $\theta_1, \theta_2, \theta_3, \theta_4$ y θ_5 compatibles con los síntomas observados. Las opiniones del equipo médico sobre estas posibilidades son tales que

$$\begin{aligned} p\{\theta_1 \cup \theta_2\} &= p\{\theta_3 \cup \theta_4 \cup \theta_5\} \\ p\{\theta_2\} &= p\{\theta_4\} = 4p\{\theta_3\} \end{aligned}$$

y creen muy remota la posibilidad de que se trate de la enfermedad θ_5 . Determinar la correspondiente distribución inicial.

Sea $p_i = p(\theta_i)$ y supongamos $p_5 = \delta$, donde $\delta > 0$ es un número pequeño, pero mayor que cero. Claramente, tenemos el sistema.

$$\begin{aligned} p_1 + p_2 &= p_3 + p_4 + \delta \\ p_2 &= p_4 \\ p_4 &= 4p_3 \\ p_1 + p_2 + p_3 + p_4 + \delta &= 1 \end{aligned}$$

(*) Si los sucesos inciertos en que estamos interesados son de la forma $\{A_i, i = 1, 2, \dots\}$ podemos considerar en su lugar la *cantidad* $\theta = 1, 2, \dots$, de forma que $p(A_i) = p(\theta = i)$.

de cuatro ecuaciones con cuatro incógnitas cuya solución, en función de δ , es $p(\theta) = \{p_1, p_2, p_3, p_4, \delta\}$ con

$$p_1 = \frac{1}{10}(1 + 8\delta), \quad p_2 = p_4 = \frac{4}{10}(1 - 2\delta), \quad p_3 = \frac{1}{10}(1 - 2\delta)$$

En particular, si se juzga 20 veces más probable que θ_1 *no* sea la causa de la dolencia a que lo sea, tendríamos $(1 - \delta)/8 = 20$, esto es $\delta = 0,048$ y por lo tanto

$$p(\theta) = \{0,138, 0,362, 0,090, 0,362, 0,048\}$$

Obsérvese que δ puede ser tan pequeño como se quiera, pero debe ser mayor que cero a menos que pueda *garantizarse* que θ_1 es *prácticamente imposible*.

Si θ es una cantidad aleatoria *continua*, su distribución de probabilidad puede ser descrita mediante la correspondiente función de densidad de probabilidad $p(\theta)$. Generalmente, se empieza eligiendo una familia de densidades de probabilidad suficientemente amplia como para contener una distribución que aproxime adecuadamente $p(\theta)$. Así, por ejemplo, si θ es una cantidad aleatoria tal que $-\infty < \theta < \infty$ y nuestra información sobre θ puede representarse por una densidad de probabilidad unimodal, simétrica y con forma acampanada, puede empezarse suponiendo que $p(\theta)$ puede aproximarse por una densidad Normal, $N(\theta|\mu, \sigma)$, de la que todavía debemos determinar sus parámetros.

Una vez elegida la familia de densidades apropiada es necesario establecer condiciones que nos permitan determinar los parámetros de la distribución, perteneciente a esa familia, que describe mejor nuestra información inicial. Estas condiciones pueden ser de muy diversos tipos, pero frecuentemente consisten en una medida de localización y una medida de dispersión de la distribución de θ . Entre las medidas de localización suele utilizarse la media o la moda; entre las de dispersión cuantiles o intervalos intercuantílicos.

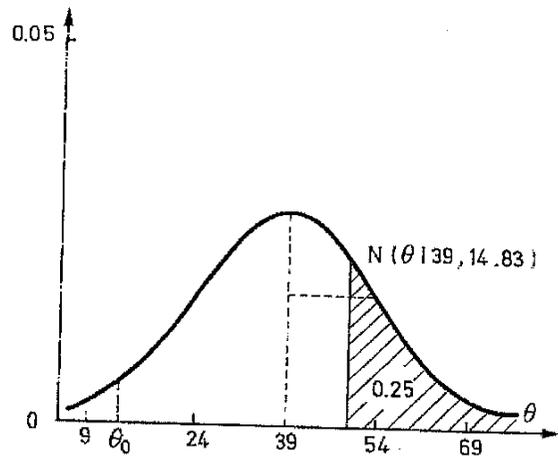
Ejemplo 5.1.2. *Cantidad de tirosina*

Las consecuencias de una determinada medicación pueden determinarse a partir de la cantidad de tirosina contenida en la orina. La información inicial sobre la cantidad de tirosina θ contenida en la orina de una determinada paciente puede describirse mediante una distribución normal centrada en 39 mg/24h y tal que $p\{\theta > 49\} = 0,25$. Determinar la correspondiente distribución inicial.

Se sabe que $p(\theta) = N(\theta|39, \sigma)$; debemos determinar σ utilizando la condición $p\{\theta > 49\} = 0,25$. Claramente, utilizando la Ecuación 4.3.1,

$$p\{\theta > 49\} = p\{49 < \theta < \infty\} = \Phi(\infty) - \Phi\left(\frac{49-39}{\sigma}\right) = 1 - \Phi\left(\frac{10}{\sigma}\right) = 0,25$$

En consecuencia, $\Phi(10/\sigma) = 0,75$; utilizando las tablas de la distribución normal, $10/\sigma = 0,6745$ y por tanto $\sigma = 14,83$. La distribución inicial es pues $N(\theta|39, 14,83)$



Una vez determinada la distribución inicial, debe procederse a calcular algunas de las probabilidades que implica; esto permite comprobar si las probabilidades calculadas son consistentes con nuestra información inicial y, consecuentemente, si la familia elegida permite una buena descripción de la información inicial.

Ejemplo 5.1.3. Cantidad de tirosoina (cont.)

Determinar el punto θ_0 tal que $p(\theta > \theta_0)/p(\theta < \theta_0) = 20$ de acuerdo con la distribución inicial obtenida en el Ejemplo 5.1.2.

Claramente,

$$\frac{p(\theta > \theta_0)}{p(\theta < \theta_0)} = \frac{p(\theta > \theta_0)}{1 - p(\theta > \theta_0)} = 20 \rightarrow p(\theta > \theta_0) = \frac{20}{21} = 0,9524$$

Utilizando de nuevo la Ecuación 4.3.1

$$p(\theta > \theta_0) = p\{\theta_0 < \theta < \infty\} = \Phi(\infty) - \Phi\left(\frac{\theta_0 - 39}{14,83}\right) = 1 - \Phi\left(\frac{\theta_0 - 39}{14,83}\right) = 0,9524$$

En las tablas de la distribución normal, puede observarse que $\Phi(1,67) = 0,9525$; por tanto,

$$\frac{\theta_0 - 39}{14,83} \approx -1,67 \rightarrow \theta_0 = 14,23$$

De forma que la distribución inicial encontrada implica que es veinte veces más probable que θ sea mayor de 14,23 mg/24 h a que θ sea menor que este valor.

Si θ es una cantidad aleatoria tal que $0 < \theta < 1$ y nuestra información sobre θ puede representarse por una densidad de probabilidad unimodal, puede empezarse suponiendo que $p(\theta)$ es una densidad Beta, $Be(\theta|\alpha, \beta)$ y precisar condiciones que permitan determinar sus parámetros.

Frecuentemente, una de estas condiciones es la *media*, m , de la distribución y la otra un *cuantil* cualquiera, esto es un valor θ_0 tal que $p[\theta \leq \theta_0] = p$, donde p es una probabilidad conocida. El valor de m es el *centro de gravedad* de la distribución; para valores de p próximos a uno, θ_0 es una especie de *cota superior*, y para valores de p próximos a cero una *cota inferior*, de los valores de θ que resultan verosímiles a la vista de la información de que se dispone.

Los parámetros α y β de la distribución inicial $Be(\theta|\alpha, \beta)$ pueden ser aproximadamente expresados en función de m , θ_0 y p , haciendo uso de la transformación normalizadora descrita en el Ejemplo 4.5.3.

Supongamos, en efecto, que $p(\theta) = Be(\theta|\alpha, \beta)$, $E(\theta) = m$ y $P(\theta \leq \theta_0) = p$. Utilizando el Teorema 4.5.3, $E(\theta) = \alpha/(\alpha + \beta)$, de forma que $m = \alpha/(\alpha + \beta)$ y, por tanto,

$$\beta = \frac{1-m}{m} \alpha \tag{1}$$

Por otra parte, si $p(\theta) = Be(\theta|\alpha, \beta)$ tenemos, utilizando los resultados del Ejemplo 4.5.3, que si $\zeta = \log\{\theta/(1-\theta)\}$,

$$p(\zeta) = N(\zeta|\mu, \sigma) \tag{2}$$

$$\mu = \log(\alpha/\beta) + (\alpha - \beta)/(2\alpha\beta)$$

$$\sigma = (\alpha + \beta)/\alpha\beta \tag{3}$$

y, en consecuencia, si $\zeta_0 = \log\{\theta_0/(1-\theta_0)\}$,

$$p = P[\theta \leq \theta_0] = P[\log\{\theta/(1-\theta)\} \leq \log\{\theta_0/(1-\theta_0)\}] = P[\zeta \leq \zeta_0] = \Phi[(\zeta_0 - \mu)/\sigma]$$

de forma que

$$\frac{\xi_0 - \mu}{\sigma} = n_p \quad (4)$$

donde n_p es la solución, que puede encontrarse en las tablas de la función de distribución normal, de la ecuación $\Phi(x) = p$.

Sustituyendo en (4) las expresiones μ y σ en función de α y β , tenemos la ecuación

$$\log [\theta_0/(1 - \theta_0)] = \log(\alpha/\beta) + (\alpha - \beta)/(2\alpha\beta) + n_p \sqrt{[(\alpha + \beta)/\alpha\beta]}$$

y utilizando (1) para poner β en función de α , resulta

$$\log \frac{\theta_0}{1 - \theta_0} = \log \frac{m}{1 - m} + \frac{2m - 1}{2(1 - m)\alpha} + \frac{n_p}{\sqrt{1 - m}} \frac{1}{\sqrt{\alpha}} \quad (5)$$

Pero (5) es una ecuación de segundo grado en $i = 1/\sqrt{\alpha}$, cuya solución es inmediata. Despejando α , resulta, para $m \neq 1/2$,

$$\alpha = \left[\frac{2a}{b + \sqrt{b^2 - 4ac}} \right]^2 \quad (6)$$

donde

$$a = (1 - 2m)/(2 - 2m)$$

$$b = n_p/\sqrt{1 - m}$$

$$c = \log \left[\frac{1 - m}{m} \frac{\theta_0}{1 - \theta_0} \right]$$

si $m < 0,5$, y estos mismos valores cambiados de signo si $m > 0,5$. Si $m = 0,5$, la ecuación (5) es de primer grado en $i = 1/\sqrt{\alpha}$; resolviéndola y despejando α , resulta

$$\alpha = 2n_p^2 / \log^2 [\theta_0/(1 - \theta_0)] \quad (7)$$

Las ecuaciones (1), (6) y (7) resuelven el problema planteado.

Una forma alternativa de especificar una distribución inicial en este problema es empezar suponiendo que $\omega = \log\{\theta/(1 - \theta)\}$ tiene *exactamente* una distribución normal, y dar condiciones que permitan determinar sus parámetros. La distribución inicial de θ será entonces *aproximadamente* Beta, y sus parámetros α y β estarán relacionados con los parámetros μ y σ de la distribución de ω mediante las ecuaciones (2) y (3).

Ejemplo 5.1.4. Tasas de morbilidad

La información de que se dispone sobre el porcentaje de personas afectadas por una epidemia en una determinada comunidad puede describirse mediante una distribución cuyo valor esperado es el 20 %; se sabe además

que la probabilidad de que tal porcentaje supere el 7 % es 0,95. Utilizando una transformación normalizadora, determinar una distribución Beta que describa aproximadamente esta información. Determinar además la probabilidad de que el porcentaje de personas afectadas por la epidemia no supere el 30 %.

Llamando θ al tanto por uno de personas afectadas, buscamos una distribución $Be(\alpha, \beta)$, cuya media sea $m = 0,2$ y tal que $p[\theta > 0,07] = 0,95$, esto es $p[\theta \leq 0,07] = 0,05$.

Con la notación que acabamos de introducir, tenemos $m = 0,2$, y $p = 0,05$; utilizando las tablas de la distribución normal (o una calculadora que disponga de esta función) encontramos $\Phi(-1,6449) = 0,05$ y, por tanto, $n_p = -1,6449$.

Utilizando las ecuaciones (6) y (1), podemos determinar los parámetros de la distribución pedida que resultan ser $\alpha = 3$ y $\beta = 12$, de forma que la distribución inicial pedida es $Be(\theta|3, 12)$.

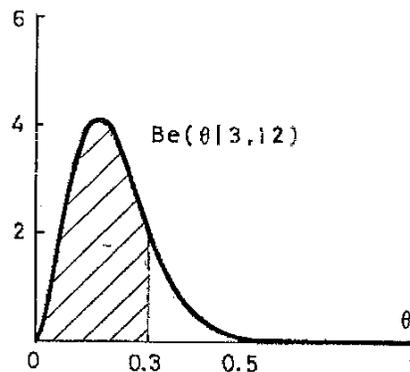
Finalmente,

$$p[\theta < 0,3] = \int_0^{0,3} Be(\theta|3, 12) d\theta$$

puede obtenerse de las tablas de la función de distribución de $Be(\theta|3, 12)$, si se dispone de ellas o, aproximadamente, mediante la transformación normalizante $\zeta = \log\{\theta/(1 - \theta)\}$. En efecto, utilizando de nuevo los resultados obtenidos en el Ejemplo 4.5.3, si θ tiene una distribución $Be(\theta|3, 12)$ entonces ζ tiene, aproximadamente, una distribución $N(\zeta| -1,51, 0,646)$. En consecuencia,

$$p[0 < \theta < 0,3] = p[-\infty < \zeta < \log(0,3/0,7)] = p[\zeta < -0,84] = \Phi\left(\frac{-0,84 + 1,51}{0,646}\right) = \Phi(1,029) \approx 0,8482$$

que no está lejos del valor exacto 0,8392 obtenido resolviendo la integral anterior por integración numérica.



Si θ es una cantidad aleatoria tal que $0 < \theta < \infty$, y nuestra información sobre θ puede representarse por una densidad de probabilidad unimodal, puede empezarse suponiendo que la distribución de θ , o la de alguna sencilla transformación suya como $1/\theta$ o $1/\theta^2$, es una distribución Gamma y especificar a continuación condiciones que permitan identificar sus parámetros. Otra alternativa es empezar suponiendo que $\log \theta$ tiene una distribución normal. Tras obtener la distribución inicial es importante comprobar, mediante el estudio de sus consecuencias, que la distribución encontrada refleja realmente la información de que dispone el decisor.

5.2. Función de verosimilitud

Frecuentemente, con objeto de mejorar nuestra información sobre el parámetro de interés θ , puede realizarse un experimento ε cuyo resultado x es una cantidad aleatoria con una distribución $p(x|\theta)$ que depende de θ ; en tal caso, la observación de x proporcionará, indirectamente, información sobre el valor de θ .

Como función de x , para un $\theta = \theta_0$ fijo, $p(x|\theta_0)$ es una densidad de probabilidad que describe la probabilidad de obtener los distintos posibles valores de x , si el parámetro de interés tuviese el valor $\theta = \theta_0$. Como función de θ , para un $x = x_0$ fijo, $p(x_0|\theta)$ es una función que describe, como vamos a ver, la verosimilitud de los distintos valores de θ a la luz del resultado experimental observado x_0 . En efecto, los valores de θ que hacen grande $p(x_0|\theta)$ son aquellos que hacían más plausible a priori la observación del resultado x_0 que ha sido eventualmente observado; en consecuencia, después de observar x_0 , estos valores de θ resultan más *verosímiles* que los demás. A la función de θ , $p(x_0|\theta)$, se le llama *función de verosimilitud* (de θ , dado x_0) y la denotaremos con $i_{x_0}(\theta)$. Debe subrayarse que la función de verosimilitud $i_{x_0}(\theta) = p(x_0|\theta)$ es una *función positiva* de θ , pero no una función de probabilidad. Consecuentemente, ni la suma de sus valores, si θ es discreta, ni su integral si es continua, tienen por qué ser la unidad.

Ejemplo 5.2.1. *Diagnosis (cont.)*

Con objeto de mejorar la información descrita en el Ejemplo 5.1.1 sobre la causa de la dolencia de un determinado paciente se realizan dos tests cuyos resultados x_1, x_2 son cantidades aleatorias X_1, X_2 cuya distribución depende de la verdadera causa de la dolencia. Específicamente, el resultado de cada uno de los tests puede ser positivo ($x_i = 1$) o negativo ($x_i = 0$) y se verifica que

$p(1, 1 \theta_1) = 0,60$	$p(1, 0 \theta_1) = 0,10$
$p(1, 1 \theta_2) = 0,10$	$p(1, 0 \theta_2) = 0,50$
$p(1, 1 \theta_3) = 0,15$	$p(1, 0 \theta_3) = 0,10$
$p(1, 1 \theta_4) = 0,10$	$p(1, 0 \theta_4) = 0,05$
$p(1, 1 \theta_5) = 0,25$	$p(1, 0 \theta_5) = 0,25$

y que $p(0, 0|\theta_i) = 0,10$ para todo θ_i . Determinar la función de verosimilitud correspondiente a los distintos resultados experimentales posibles.

Si los dos tests dan resultados positivos, esto es si se observa $(1, 1)$, la función de verosimilitud correspondiente es de acuerdo con la tabla

$$i_{1,1}(\theta) = \{0,60, 0,10, 0,15, 0,10, 0,25\}$$

lo que aumenta la verosimilitud de θ como causa de la dolencia. Obsérvese que $i_{1,1}(\theta)$ no es una distribución de probabilidad de θ ; sus elementos no suman la unidad.

Análogamente, si se observa $(1, 0)$, la función de verosimilitud correspondiente es

$$i_{1,0}(\theta) = \{0,10, 0,50, 0,10, 0,05, 0,25\}$$

de forma que el resultado $(1, 0)$ aumenta la verosimilitud de θ_1 , y en parte la de θ_5 , como posibles causas de la dolencia.

Puesto que para θ_i fijo, $p(x_1, x_2|\theta_i)$ debe ser una distribución de probabilidad y por lo tanto sumar la unidad, las probabilidades correspondientes a que el primer test de negativo y el segundo positivo son

$$\begin{aligned} p(0, 1|\theta_1) &= 0,20 \\ p(0, 1|\theta_2) &= 0,30 \\ p(0, 1|\theta_3) &= 0,65 \\ p(0, 1|\theta_4) &= 0,75 \\ p(0, 1|\theta_5) &= 0,40 \end{aligned}$$

de forma que la función de verosimilitud correspondiente al resultado $(0, 1)$ es

$$i_{0,1}(\theta) = \{0,20, 0,30, 0,65, 0,75, 0,40\}$$

Finalmente, la función de verosimilitud correspondiente a $(0, 0)$ es

$$i_{0,0}(\theta) = \{0,10, 0,10, 0,10, 0,10, 0,10\}$$

esto es uniforme, de forma que, en este ejemplo, si ambos tests dan resultados negativos no obtenemos información alguna sobre la verdadera causa de la dolencia.

Consideraremos, a continuación, dos de los tipos más frecuentes de función de verosimilitud: la que se obtiene de la observación de sucesos de Bernoulli y la que aparece al realizar medidas normales.

Supóngase, por ejemplo, que se está interesado en la morbilidad de una determinada enfermedad, esto es en la frecuencia de su aparición en una determinada población; si se denota con θ la probabilidad de que una persona cualquiera de la población tenga la enfermedad, estamos interesados en el valor de θ . La información de que inicialmente disponemos sobre el valor de θ vendrá descrita, de acuerdo con lo expuesto en la sección anterior por una determinada distribución inicial $p(\theta)$. Una forma de mejorar esta información es observar un elemento de la población y comprobar si tiene o no la enfermedad.

Si denotamos por x el resultado de una observación, con $x = 1$, si tiene la enfermedad y $x = 0$ si no la tiene, esto es $p(x|\theta) = Br(x|\theta)$ de forma que $l_1(\theta) = \theta$ es la función de verosimilitud de $x = 1$ y $l_0(\theta) = 1 - \theta$ la de $x = 0$.

DEFINICIÓN 5.2.1. Llamaremos muestra aleatoria de tamaño n de una población cuyos elementos tienen una distribución $p(x|\theta)$ que depende de θ , a un conjunto (x_1, x_2, \dots, x_n) de observaciones independientes dado θ , esto es tales que $p(x_1, x_2, \dots, x_n|\theta) = p(x_1|\theta) p(x_2|\theta) \dots p(x_n|\theta)$.

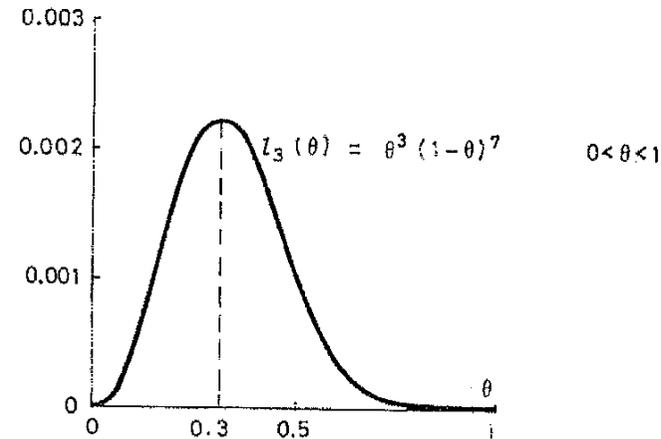
Supóngase ahora que se extrae una muestra aleatoria de tamaño n de la población que queremos analizar, esto es se observan n elementos de la población escogidos al azar y se examina cada uno de ellos para determinar si tiene o no la enfermedad. Denotando por x_i el resultado correspondiente ($x_i = 1$ si el elemento i tiene la enfermedad, $x_i = 0$ si no la tiene), de forma que para todo x_i , $p(x_i|\theta) = Br(x_i|\theta)$, tendremos

$$p(x_1, \dots, x_n|\theta) = p(x_1|\theta) \dots p(x_n|\theta) = \theta^r (1 - \theta)^{n-r}$$

donde $r = \sum x_i$ es el número de elementos en la muestra de n que tenían la enfermedad. En consecuencia,

$$l_3(\theta) = \theta^r (1 - \theta)^{n-r} \quad (1)$$

es la función de verosimilitud correspondiente a la obtención de r enfermos en una muestra aleatoria de n individuos. Por ejemplo, para $n = 10$ y $r = 3$ obtenemos $l_3(\theta) = \theta^3 (1 - \theta)^7$.



La observación de la representación gráfica correspondiente hace patente que, a la vista de esos resultados experimentales, los valores de θ alrededor de 0,3 (en general de r/n) resultan más verosímiles que los demás.

DEFINICIÓN 5.2.2. Llamaremos estimador máximo-verosímil de θ al valor $\hat{\theta}$ de θ que maximiza la función de verosimilitud.

En el caso anterior, el estimador máximo-verosímil es el valor de θ que maximiza (1), esto es $\hat{\theta} = r/n$.

Consideremos otra situación frecuente; supóngase que se está interesado en la temperatura μ de un determinado paciente. Claramente, se dispone de cierta información inicial sobre el valor de μ ; se sabe que está necesariamente entre 35 y 42 °C, que su valor esperado para una persona sana es 36,5 °C, etc.; esta información estará contenida en la distribución inicial de μ , $p(\mu)$. Si necesitamos una información más precisa recurrimos a un experimento obvio: medir con un termómetro la temperatura del paciente.

Toda medida está sujeta a distintos tipos de error. Si las causas de estos posibles errores son numerosas, todas del mismo orden de magnitud, e independientes, el teorema central del límite (Sección 4.3) permite asegurar que el valor medido es una cantidad aleatoria con distribución aproximadamente normal centrada en la medida verdadera. En nuestro caso, podemos suponer pues que la lectura dada por el termómetro es una cantidad aleatoria x con distribución normal $N(x|\mu, \sigma)$ centrada en la temperatura verdadera del paciente μ y desviación típica σ °C que depende de la construcción del termó-

metro. Si se nos dice, por ejemplo, que $\sigma = 0,05$, la función de verosimilitud correspondiente a una lectura x del termómetro será

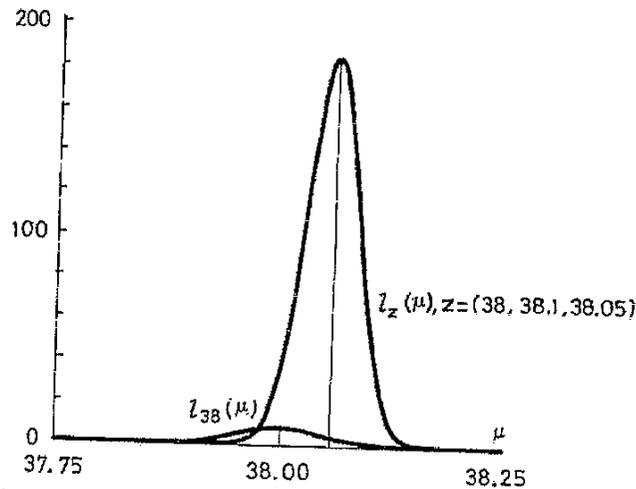
$$l_x(\mu) = \frac{1}{0,05\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{0,05} \right)^2 \right\}$$

la observación de la gráfica obtenida para una lectura $x = 38^\circ\text{C}$ permite reconocer que, después del experimento, todos los valores de μ alejados de 38°C resultan claramente *imverosímiles*.

El conjunto de valores verosímiles de μ puede todavía reducirse realizando nuevas medidas. Así, si medimos n veces consecutivas la temperatura del paciente, que suponemos constante en ese tiempo, obteniendo las lecturas $z = \{x_1, x_2, \dots, x_n\}$, la correspondiente función de verosimilitud será

$$l_z(\mu) = \left(\frac{1}{0,05\sqrt{2\pi}} \right)^n \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{0,05} \right)^2 \right\}$$

la observación de la gráfica obtenida para tres lecturas $38, 38,1$ y $38,05^\circ\text{C}$ permite reconocer que después de ellas, todavía es más reducido el conjunto de valores verosímiles de μ .



Obsérvese que, en las funciones de verosimilitud, lo único que importan son las alturas relativas para una *misma* curva; no son distribuciones de probabilidad y, por tanto, no tienen por que integrar la unidad.

En general, la función de verosimilitud correspondiente a una muestra aleatoria $\{x_1, x_2, \dots, x_n\}$ extraída de una población normal $N(x|\mu, \sigma)$ será de la forma

$$\begin{aligned} p(x_1, \dots, x_n|\mu, \sigma) &= \prod_{i=1}^n N(x_i|\mu, \sigma) = \\ &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} = \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} \end{aligned} \quad (2)$$

Como veremos más adelante esta expresión juega un papel central en la deducción de una parte importante de los resultados de uso más frecuentes en estadística. Es fácil comprobar que, como función de μ , la expresión (2) alcanza su máximo para $\bar{x} = (\sum x_i)/n$ y como función de σ^2 para $s^2 = \sum (x_i - \bar{x})^2/n$ que son, por tanto, los correspondientes *estimadores máximo-verosímiles*.

5.3. Distribución predictiva

Considérese la distribución $p(x|\theta)$ del resultado x de un determinado experimento ε . Generalmente, el valor de θ es desconocido y, en consecuencia, $p(x|\theta)$ no puede utilizarse para determinar los valores de x que resultan más probables. Sin embargo, aunque el valor exacto de θ es desconocido, siempre se dispone de cierta información sobre θ , que es la descrita por su distribución inicial $p(\theta)$. Esta información puede combinarse con la distribución $p(x|\theta)$ para poder describir la información que se posee sobre los posibles valores de x . En efecto,

$$p(x) = \sum p(x|\theta_i) p(\theta_i) \quad (1)$$

si θ es una cantidad aleatoria discreta y

$$p(x) = \int p(x|\theta) p(\theta) d\theta \quad (2)$$

si se trata de una cantidad aleatoria continua, proporcionan una distribución de x totalmente conocida que podemos utilizar, entre otras cosas, para hacer *predicciones* sobre los valores de x a que dará lugar el experimento ε , por lo que recibe el nombre de *distribución predictiva*. Naturalmente, en las expre-

siones (1) y (2), $p(x)$ es una función de probabilidad o una función de densidad de probabilidad según x sea una cantidad aleatoria discreta o una continua.

Puede observarse que la expresión (1) no es más que una forma del teorema de la probabilidad total (Teorema 3.4.1), y que (2) no es más que la forma continua de (1) en la que la función de probabilidad $p(\theta_i)$ se sustituye por la de densidad de probabilidad $p(\theta)$ y la operación suma por la operación integral.

Ejemplo 5.3.1. *Diagnosis (cont.)*

Determinar la distribución predictiva de los resultados de los tests descritos en el Ejemplo 5.2.1 utilizando la distribución inicial sobre las posibles causas de la dolencia construida en el Ejemplo 5.1.1.

De acuerdo con la ecuación (1), la distribución predictiva del resultado (x_1, x_2) de ambos tests vendrá dada por

$$p(x_1, x_2) = \sum_{i=1}^2 p(x_1, x_2 | \theta_i) p(\theta_i), \quad x_1 = 0, 1, \quad x_2 = 0, 1$$

En el Ejemplo 5.1.1 se determinó la distribución inicial $p(\theta) = \{0.138, 0.362, 0.090, 0.362, 0.048\}$ y en el Ejemplo 5.2.1 se especifican los valores de $p(x_1, x_2 | \theta_i)$. Operando, resulta

$$p(1, 1) = 0.60 \times 0.138 + 0.10 \times 0.362 + 0.15 \times 0.090 + \\ \times 0.10 \times 0.362 + 0.25 \times 0.048 = 0.181$$

Análogamente,

$$p(1, 0) = 0.235; \quad p(0, 1) = 0.484; \quad p(0, 0) = 0.100$$

Naturalmente, la suma de los cuatro valores obtenidos es la unidad, puesto que se trata de una distribución de probabilidad. Resulta, pues, que el resultado más probable del experimento, con probabilidad 0.484, es (0, 1), esto es que el primer test de negativo y el segundo positivo.

Las correspondientes distribuciones marginales son

$$\begin{cases} p(x_1 = 1) = 0.181 + 0.235 = 0.416 \\ p(x_1 = 0) = 0.484 + 0.100 = 0.584 \\ p(x_2 = 1) = 0.181 + 0.484 = 0.665 \\ p(x_2 = 0) = 0.235 + 0.100 = 0.335 \end{cases}$$

de forma que se tiene probabilidad 0.584 de que el primer test de negativo y 0.665 de que el segundo de positivo.

Consideraremos ahora distribuciones predictivas correspondientes a las situaciones experimentales mencionadas en la sección anterior.

Supóngase por ejemplo que piensa extraerse una muestra aleatoria de tamaño n , $\{x_1, x_2, \dots, x_n\}$ de una determinada población con objeto de determinar en cada uno de los elementos de la muestra si tienen ($x_i = 1$) o no tienen ($x_i = 0$) una determinada característica, y precisar el número $r = \sum x_i$ de los que la tienen. Obviamente, r puede tomar cualquiera de los valores 0, 1, 2, ..., n . Si supiésemos cual es la probabilidad θ de que un elemento cualquiera tenga la característica objeto de la investigación, conoceríamos la distribución de r . En efecto si $p(x_i | \theta) = \theta$ entonces, según vimos en la Sección 4.2,

$$p(r | \theta, n) = B_b(r | \theta, n) = \binom{n}{r} \theta^r (1 - \theta)^{n-r} \quad (3)$$

Si el verdadero valor de θ es desconocido, será necesario limitarse a utilizar la información inicial que se posea sobre su valor, información que vendrá descrita por su distribución inicial $p(\theta)$. En tal caso, utilizando (2), la distribución *predictiva* de r vendrá dada por

$$p(r | n) = \int_0^1 p(r | \theta, n) p(\theta) d\theta \quad (4)$$

Frecuentemente, la información inicial sobre θ puede ser aproximadamente descrita por una distribución de la familia Beta. En tal caso, si $p(\theta) = Be(\theta | \alpha, \beta)$ la expresión (4) resulta ser

$$\begin{aligned} p(r | n) &= \int_0^1 B_b(r | \theta, n) Be(\theta | \alpha, \beta) d\theta \\ &= \binom{n}{r} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \int_0^1 \theta^{r+\alpha-1} (1 - \theta)^{n-r+\beta-1} d\theta = \\ &= \binom{n}{r} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \frac{\Gamma(\alpha + r) \Gamma(\beta + n - r)}{\Gamma(\alpha + \beta + n)} \end{aligned} \quad (5)$$

DEFINICIÓN 5.3.1. Una cantidad aleatoria discreta x tiene una distribución Beta-binomial si su función de probabilidad, que denotaremos $Bb(x | \alpha, \beta, n)$ es de la forma

$$\begin{aligned} p(x) &= Bb(x | \alpha, \beta, n) = \\ &= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \frac{\Gamma(\alpha + x) \Gamma(\beta + n - x)}{\Gamma(\alpha + \beta + n)}, \quad x = 0, 1, \dots, n \\ &= 0, \quad \text{para cualquier otro } x. \end{aligned}$$

Utilizando la Definición 5.3.1, la ecuación (5) puede reescribirse en la forma

$$\int_0^1 Bb(r|\theta, n) Be(\theta|\alpha, \beta) d\theta = Bb(r|\alpha, \beta, n) \quad (6)$$

Ejemplo 5.3.2. Tasas de morbilidad (cont.)

Con objeto de mejorar la información disponible sobre la incidencia de una epidemia de gripe, un equipo se dispone a obtener una muestra de 10 elementos de la población aleatoriamente elegidos. Si la información de que el equipo dispone sobre la probabilidad θ de que un individuo cualquiera tenga gripe puede describirse mediante la distribución $Be(\theta|3, 12)$, construida en el Ejemplo 5.1.4, determinar la distribución predictiva del número r de individuos entre esos 10 que los investigadores pueden esperar que tengan gripe.

De acuerdo con la ecuación (6), la distribución pedida es la Beta-binomial $Bb(r|3, 12, 10)$; en consecuencia,

$$p(r) = \binom{10}{r} \frac{\Gamma(15)}{\Gamma(3)\Gamma(12)} \frac{\Gamma(3+r)\Gamma(12+10-r)}{\Gamma(15+10)}$$

Dando valores, se obtiene la distribución

r	$p(r) = Bb(r 3, 12, 10)$
0	0,1798
1	0,2569
2	0,2312
3	0,1623
4	0,0947
5	0,0468
6	0,0195
7	0,0067
8	0,0018
9	0,0003
10	0,00003

de manera que, en particular, lo más probable, es que aparezcan uno o dos elementos con gripe (probabilidades 0,2569 y 0,2312 respectivamente), aunque 0 y 3 son asimismo

valores bastante probables. Valores de r superiores a 5 son sin embargo muy poco probables.

Supongamos ahora que desea medirse una magnitud cuyo verdadero valor, desconocido, denotaremos μ , y que la medida que va a obtenerse x tiene una distribución normal centrada en μ y con una desviación típica conocida σ . Si conociésemos μ , conoceríamos totalmente la distribución de x que sería

$$p(x|\mu) = N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$

Como el valor de μ es desconocido, por eso efectuamos la medida x , tendremos que limitarnos a utilizar la información que poseemos sobre su valor, descrita por la distribución inicial $p(\mu)$.

En tal caso, utilizando (2), la distribución *predictiva* de x vendrá dada por

$$p(x) = \int_{-\infty}^{+\infty} N(x|\mu, \sigma) p(\mu) d\mu \quad (7)$$

Frecuentemente, la información inicial sobre μ puede describirse mediante una distribución de la familia Normal. En tal caso, si $p(\mu) = N(\mu|\mu_0, \sigma_0)$, la expresión (7) resulta ser

$$p(x) = \int_{-\infty}^{+\infty} N(x|\mu, \sigma) N(\mu|\mu_0, \sigma_0) d\mu = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} \times \frac{1}{\sigma_0\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{\mu-\mu_0}{\sigma_0} \right)^2 \right\} d\mu$$

Resolviendo la integral y reordenando el resultado, resulta

$$p(x) = \frac{1}{\sqrt{(\sigma_0^2 + \sigma^2)}\sqrt{(2\pi)}} \exp \left\{ -\frac{1}{2} \frac{(x-\mu_0)^2}{\sigma_0^2 + \sigma^2} \right\}$$

que es una densidad normal de media μ_0 y varianza $\sigma_0^2 + \sigma^2$. Este resultado puede escribirse en la forma

$$\int N(x|\mu, \sigma) \cdot N(\mu|\mu_0, \sigma_0) d\mu = N(x|\mu_0, \sqrt{(\sigma^2 + \sigma_0^2)}) \quad (8)$$

Ejemplo 5.3.3. Cantidad de tirosina (cont.)

Con objeto de adquirir información sobre las condiciones en que se encuentra un determinado paciente, va a medirse la cantidad de tirosina (en mg/24 h) contenida en su orina. Debido a distintos errores experimentales, el valor obtenido no será, en general, el verdadero valor μ , sino una cantidad aleatoria x con una distribución normal centrada en μ y con una desviación típica σ que depende del proceso de medida.

Supongamos que otras medidas realizadas con el mismo aparato permiten suponer que $\sigma = 2$ mg/24 h. Si la información inicial sobre el valor de μ puede describirse mediante la distribución $N(\mu|39, 14,83)$ construida en el Ejemplo 5.1.2, determinar la distribución predictiva del valor x que va a ser obtenido.

La distribución condicional de x es $N(x|\mu, 2)$ y la distribución inicial de θ es $N(\mu|39, 14,83)$. En consecuencia, de acuerdo con (8), la distribución predictiva de x será normal con media 39 y desviación típica $\sqrt{14,83^2 + 2^2} = 14,96$.

Por ejemplo, es poco probable obtener valores de x superiores a 60 puesto que

$$p[x > 60] = p[60 < x < \infty] = \Phi(\infty) - \Phi\left(\frac{60 - 39}{14,96}\right) = 1 - \Phi(1,404) \approx 0,080$$

y muy probable que sean mayores que 10 puesto que, como se comprueba con facilidad, $p[x > 10] \approx 0,974$.

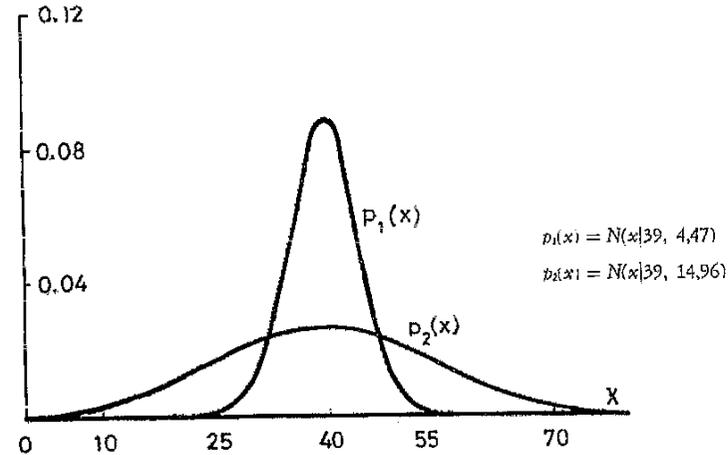
La distribución predictiva $p(x)$ juega, respecto al valor experimental x que va a ser obtenido, el mismo papel que la distribución inicial $p(\theta)$ respecto al valor θ del parámetro de interés: ambas distribuciones describen la información de que se dispone; sobre el valor x que va a observarse en el caso de la distribución predictiva, sobre el verdadero valor de θ en el de la distribución inicial. La ecuación que las relaciona, $p(x) = \int p(x|\theta)p(\theta)d\theta$, permite analizar las consecuencias de adoptar una determinada distribución inicial $p(\theta)$ en términos de sus implicaciones sobre la plausibilidad de los distintos valores posibles de x , descrita por $p(x)$.

Ejemplo 5.3.4. Cantidad de tirosina (cont.)

Con los mismos supuestos del Ejemplo 5.3.3, supongamos que la información inicial sobre μ puede describirse mediante la distribución $N(\mu|39, 4)$. Determinar la nueva distribución predictiva.

La distribución condicional sigue siendo $N(x|\mu, 2)$, mientras la inicial es ahora $N(\mu|39, 4)$ esto es con la misma media pero menor desviación típica. De acuerdo con (8)

la distribución predictiva de x será ahora normal de media 39 y desviación típica $\sqrt{4^2 + 2^2} = 4,47$.



con lo que $p[x > 60] \approx 0$ y $p[x > 10] \approx 1$.

La mayor información inicial sobre μ que implica la reducción en la desviación típica de su distribución inicial se traduce en una mayor información sobre los valores de x que pueden ser esperados. Así, la mayor información inicial de que se dispone en el segundo caso permite reducir el rango de valores plausibles de x , esto es el conjunto de valores de x con densidad de probabilidad apreciablemente positiva, de (3, 75) a (27, 51).

5.4. Teorema de Bayes y distribución final

Supongamos que, con objeto de mejorar la información de que se dispone sobre el valor de θ , se realiza un experimento e que da lugar a unos resultados x con probabilidad (o densidad de probabilidad si x es una cantidad aleatoria continua) $p(x|\theta)$. Como función de x y θ , $p(x|\theta)$ es el *modelo probabilístico* que define la relación que se supone entre los *resultados experimentales* x y el *parámetro de interés* θ . Sea $p(\theta)$ la distribución inicial de θ . Después de observar el resultado x del experimento, la información de que disponemos sobre el valor de θ estará descrita por su *distribución final* $p(\theta|x)$. El *Teorema de Bayes* permite obtener la distribución final $p(\theta|x)$ a partir de la distribución inicial $p(\theta)$ y de la función de verosimilitud del resultado obtenido $p(x|\theta)$.

TEOREMA 5.4.1 (Teorema de Bayes). Sean x los resultados del experimento ϵ definido mediante el modelo $p(x|\theta)$ y sea $p(\theta)$ la distribución inicial de θ . La distribución final de θ es entonces

$$p(\theta|x) = p(x|\theta) p(\theta)/p(x) \quad (1)$$

donde $p(x)$ es la distribución predictiva de x .

El teorema anterior es una generalización natural a funciones de probabilidad, o de densidad de probabilidad, del Teorema de Bayes para sucesos estudiado en la Sección 3.4.

Sin pérdida de generalidad, el Teorema de Bayes puede expresarse en la forma

$$p(\theta|x) \propto p(x|\theta) p(\theta) = I_x(\theta) p(\theta) \quad (2)$$

de manera que si dos modelos probabilísticos son proporcionales para todo θ , dan necesariamente lugar a la misma distribución final, y puede escribirse

Distribución final \propto Verosimilitud \times Distribución inicial

En efecto, la constante de proporcionalidad puede determinarse en cualquier momento utilizando la propiedad de que, por ser $p(\theta|x)$ una distribución de probabilidad, $\int p(\theta|x) d\theta = 1$ en el caso continuo, y $\sum p(\theta_i|x) = 1$ en el caso discreto. Específicamente, si $p(\theta|x) = C p(x|\theta) p(\theta)$

$$\begin{aligned} \int p(\theta|x) d\theta &= C \int p(x|\theta) p(\theta) d\theta = 1 \\ C &= 1 / \int p(x|\theta) p(\theta) d\theta = 1/p(x) \end{aligned} \quad (3)$$

y, análogamente,

$$\begin{aligned} \sum p(\theta_i|x) &= C \sum p(x|\theta_i) p(\theta_i) = 1 \\ C &= 1 / \sum p(x|\theta_i) p(\theta_i) = 1/p(x) \end{aligned} \quad (4)$$

la expresión (2) es generalmente más cómoda de manejar que la expresión (1).

La distribución final $p(\theta|x)$ combina la información inicial sobre θ , contenida en $p(\theta)$ con la información sobre θ proporcionada por el resultado experimental x . La distancia entre $p(\theta|x)$ y $p(\theta)$ es pues una medida de la información proporcionada por x sobre el valor de θ .

DEFINICIÓN 5.4.1. Sean x los resultados de un experimento ϵ cuya distribución es $p(x|\theta)$ y sea $p(\theta)$ la distribución inicial de θ . La información proporcionada por x sobre el valor de θ es entonces

$$I^0\{\epsilon, p(\theta)|x\} = \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} d\theta \quad (5)$$

en el caso continuo y

$$I^0\{\epsilon, p(\theta)|x\} = \sum p(\theta_i|x) \log \frac{p(\theta_i|x)}{p(\theta_i)} \quad (6)$$

en el discreto.

Puede observarse que la información proporcionada por x es el valor esperado, después de haber observado x , de la diferencia $\log p(\theta|x) - \log p(\theta)$, esto es de la diferencia entre los logaritmos de las probabilidades (densidades de probabilidad) asociadas al verdadero valor de θ después y antes del experimento. Es fácil demostrar que para cualquier experimento y cualquier distribución inicial $I^0\{\epsilon, p(\theta)|x\}$ no es nunca menor que cero, esto es los datos no pueden proporcionar información negativa, y es cero si, y solamente si, $p(\theta|x) = p(\theta)$ esto es si, y solamente si, los datos no modifican la distribución inicial. Razonamientos más completos que justifican esta definición de información pueden encontrarse en Good (1966) y Bernardo (1979 a).

TEOREMA 5.4.2. El valor esperado de la cantidad de información necesaria verdadero valor de una cantidad aleatoria discreta θ con función de probabilidad $p(\theta)$ viene dado por

$$H\{p(\theta)\} = - \sum p(\theta_i) \log p(\theta_i) \quad (7)$$

Demostración

Para determinar el verdadero valor de θ necesitaríamos unos resultados x tales que la correspondiente distribución final asignase probabilidad 1 al verdadero valor de θ y cero a todos los demás. Utilizando la Definición 5.4.1 para esta distribución final, y teniendo en cuenta que

$$\lim_{x \rightarrow \theta} x \log(x) = 0$$

obtenemos (7) al considerar valores esperados.

Por motivos históricos, relacionados con su aparición en termodinámica, la expresión $H\{p(\theta)\}$ recibe el nombre de *entropía* de la distribución $p(\theta)$.

DEFINICIÓN 5.4.2. La cantidad de información sobre θ que puede esperarse de los resultados x de un experimento ϵ viene dada por

$$I^{\theta}\{\epsilon, p(\theta)\} = \int p(x) I^{\theta}\{\epsilon, p(\theta|x)\} dx \quad (8)$$

si x es una cantidad aleatoria continua o bien por

$$I^{\theta}\{\epsilon, p(\theta)\} = \sum p(x_i) I^{\theta}\{\epsilon, p(\theta|x_i)\} \quad (9)$$

si x es discreta.

Cuando se utiliza la base dos para los logaritmos, las unidades de información obtenidas se llaman *bits* (*) y describen el número de preguntas binarias sobre θ con respuestas equiprobables cuyas respuestas proporcionarían, en valor medio, la misma información (Renyi, 1962/1966, pág. 564). Naturalmente, basta dividir por $\log(2) \approx 0,69315$ los resultados de operar con logaritmos neperianos para obtener el resultado en bits.

Ejemplo 5.4.1. Diagnósis (cont.)

Las consecuencias de un determinado tratamiento dependen de la enfermedad del paciente, que puede ser $\theta_1, \theta_2, \theta_3, \theta_4$ o θ_5 . La opinión inicial del equipo médico puede describirse mediante la distribución inicial.

$$p(\theta) = \{0,138, 0,362, 0,090, 0,362, 0,048\}$$

construida en el Ejemplo 5.1.1. Con objeto de mejorar esta información se realizan los tests descritos en el Ejemplo 5.2.1. Determinar la distribución final correspondiente a cada uno de los resultados posibles, y la cantidad de información sobre θ que estos resultados proporcionan.

De acuerdo con los datos del Ejemplo 5.2.1, la tabla de funciones de verosimilitud es

	(1, 1)	(1, 0)	(0, 1)	(0, 0)
θ_1	0,60	0,10	0,20	0,10
θ_2	0,10	0,50	0,30	0,10
θ_3	0,15	0,10	0,65	0,10
θ_4	0,10	0,05	0,75	0,10
θ_5	0,25	0,25	0,40	0,10

(*) El término *bit* es una contracción de la expresión inglesa *binary digit*, dígito binario. En efecto, en un número escrito en base 2, cada dígito proporciona precisamente una unidad, en base 2, de información puesto que sólo puede tomar dos valores, 0 o 1, equiprobables $v \rightarrow -0,5 \log_2(0,5) = 0,5 \log_2(0,5) = 1$.

de forma que, por ejemplo, $p(1, 1|\theta_1) = 0,10$. Si el resultado del experimento es (1, 1), la distribución final será de la forma

$$\begin{aligned} p(\theta_1|1, 1) &\propto p(1, 1|\theta_1) p(\theta_1) = 0,60 \times 0,138 = 0,083 \\ p(\theta_2|1, 1) &\propto p(1, 1|\theta_2) p(\theta_2) = 0,10 \times 0,362 = 0,036 \\ p(\theta_3|1, 1) &\propto p(1, 1|\theta_3) p(\theta_3) = 0,15 \times 0,090 = 0,014 \\ p(\theta_4|1, 1) &\propto p(1, 1|\theta_4) p(\theta_4) = 0,10 \times 0,362 = 0,036 \\ p(\theta_5|1, 1) &\propto p(1, 1|\theta_5) p(\theta_5) = 0,25 \times 0,048 = 0,012 \end{aligned}$$

$$0,181 = p(1, 1)$$

En virtud de (4), la suma de los valores obtenidos, 0,181 es $p(1, 1)$, lo que coincide con el valor encontrado en el Ejemplo 5.3.1, y la constante de proporcionalidad $1/p(1, 1) = 1/0,181 = 5,525$ con lo que la distribución final de θ es

$$p(\theta|1, 1) = \{0,459, 0,199, 0,077, 0,199, 0,066\}$$

Comparando este resultado con la distribución inicial se observa que, como era de esperar, la probabilidad asignada a θ_1 se ha incrementado mucho como consecuencia del resultado experimental, pasando de 0,138 a 0,459, mientras las de θ_2 y θ_4 disminuyen fuertemente.

De forma totalmente análoga, puede obtenerse que

$$\begin{aligned} p(\theta|1, 0) &= \{0,060, 0,771, 0,039, 0,078, 0,052\} \\ p(\theta|0, 1) &= \{0,057, 0,224, 0,122, 0,562, 0,035\} \\ p(\theta|0, 0) &= \{0,138, 0,362, 0,090, 0,362, 0,048\} \end{aligned}$$

Como era de esperar, puesto que $p(\theta, 0|\theta_i)$ es constante para todos los θ_i , el resultado (0, 0) no proporciona información alguna sobre θ y resulta que $p(\theta|0, 0) = p(\theta)$, esto es la distribución final es igual a la inicial.

La información sobre θ proporcionada por cada uno de los resultados experimentales es, utilizando (6),

$$\begin{aligned} I^{\theta}\{\epsilon, p(\theta)|(1, 1)\} &= 0,3225 \\ I^{\theta}\{\epsilon, p(\theta)|(1, 0)\} &= 0,4606 \\ I^{\theta}\{\epsilon, p(\theta)|(0, 1)\} &= 0,5541 \\ I^{\theta}\{\epsilon, p(\theta)|(0, 0)\} &= 0 \end{aligned}$$

de manera que el resultado más informativo, el que más cambia las opiniones iniciales, es (0, 1) mientras que (0, 0) no proporciona información alguna porque no modifica las opiniones iniciales. La información que podrá esperarse sobre θ antes de observar x será, según (8) y los resultados del Ejemplo 5.3.1 de 0,4348, esto es de 0,6273 bits, comparable por tanto a un 62% de la que proporcionaría una pregunta binaria sobre el verdadero valor de θ con sus dos respuestas equiprobables.

Consideremos ahora las distribuciones finales correspondientes a las situaciones experimentales descritas en la Sección 5.2.

Supongamos primero que, con objeto de mejorar nuestra información inicial sobre la proporción θ de elementos de una población que poseen una determinada característica se observa una muestra aleatoria de n elementos, r de los cuales resultan poseer la característica investigada. Si la información inicial sobre θ es descrita por la distribución inicial $p(\theta)$, la correspondiente distribución final será, en virtud de 5.2.1 y del Teorema de Bayes, de la forma

$$p(\theta|r, n) \propto p(r|n, \theta) p(\theta) \\ \propto \theta^r (1 - \theta)^{n-r} p(\theta)$$

Si además, la distribución inicial es de la familia Beta, $p(\theta) = Be(\theta|\alpha, \beta)$ esto es, según la definición 4.3.3, de la forma, $p(\theta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}$, resulta

$$p(\theta|r, n) \propto \theta^{\alpha+r-1} (1 - \theta)^{\beta+n-r-1}$$

y, por tanto, comparando con la definición de distribución Beta,

$$p(\theta|r, n) = Be(\theta|\alpha + r, \beta + n - r) \quad (10)$$

En este tipo de cálculos puede observarse la ventaja de utilizar el Teorema de Bayes en su forma proporcional dada por (2).

Ejemplo 5.4.2. Tasa de morbilidad (cont.)

La información sobre la proporción θ de personas afectadas por una epidemia de gripe puede describirse mediante la distribución $Be(\theta|3, 12)$ construida en el Ejemplo 5.1.4. Para mejorar esta información, se observan 10 individuos extraídos al azar, de los que 2 estaban afectados. Determinar la correspondiente distribución final y , utilizando una aproximación normal, la probabilidad final de que $\theta < 0,3$.

De acuerdo con la ecuación (10), la distribución final es Beta con parámetros $3 + 2 = 5$ y $12 + 10 - 2 = 20$. La probabilidad pedida es

$$p[\theta < 0,3 | r = 2, n = 10] = \int_0^{0,3} Be(\theta|5, 20) d\theta$$

cuyo cálculo exacto es complicado. Sin embargo, utilizando los resultados del Ejemplo 4.5.3, si θ tiene una distribución $Be(\theta|5, 20)$, $\zeta = \log\{\theta/(1 - \theta)\}$ tiene una distribución aproximadamente Normal $N(\zeta|-1,461, 0,5)$ y por tanto

$$p[\theta < 0,3 | r = 2, n = 10] = p\left[\zeta < \log \frac{0,3}{0,7}\right] = p[-\infty < \zeta < -0,847] = \\ = \Phi\left(\frac{-0,847 + 1,461}{0,5}\right) = \Phi(1,288) = 0,8903$$

Comparando con el valor 0,8482 obtenido en el Ejemplo 5.1.4 para la probabilidad del mismo intervalo en la distribución inicial, observamos que el resultado obtenido (2 enfermos entre 10 personas), aumenta esta probabilidad, lo que resulta muy razonable puesto que este resultado aumenta la verosimilitud de los valores de θ próximos a la proporción muestral $r/n = 0,2$.

El Teorema de Bayes puede ser utilizado de forma iterativa; en efecto, si tras observar los resultados experimentales x y transformar $p(\theta)$ en $p(\theta|x)$ deseamos realizar un segundo experimento cuyo resultado denotaremos por y , podemos obtener la nueva distribución final $p(\theta|x, y)$ utilizando $p(\theta|x)$ como distribución inicial. El resultado,

$$p(\theta|x, y) \propto p(y|\theta) p(\theta|x) \quad (11)$$

es naturalmente el mismo que el que habríamos obtenido analizando simultáneamente ambos resultados experimentales; en efecto, en este último caso, tendríamos

$$p(\theta|x, y) \propto p(x, y|\theta) p(\theta) = p(x|\theta) p(y|\theta) p(\theta) \\ \propto p(y|\theta) p(x|\theta) p(\theta)/p(x) = p(y|\theta) p(\theta|x)$$

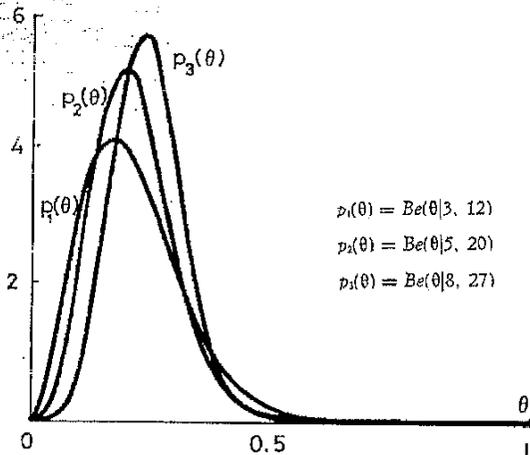
que es la expresión (11).

Ejemplo 5.4.3. Tasa de morbilidad (cont.)

Supongamos que con objeto de obtener todavía una información más precisa sobre la proporción θ de enfermos de gripe, se obtiene una nueva muestra de 10 elementos, 3 de los cuales resultan estar afectados. Determinar la nueva distribución final.

La nueva distribución inicial es la distribución final obtenida en el Ejemplo 5.4.2, esto es la distribución $Be(\theta|5, 20)$. De acuerdo con (10) la nueva distribución final será Beta con parámetros $5 + 3 = 8$ y $20 + 10 - 3 = 27$, esto es $Be(\theta|8, 27)$.

El mismo resultado se puede obtener combinando las dos muestras $r = 2 + 3 = 5$, $n = 10 + 10 = 20$ y utilizando la distribución inicial original $Be(\theta|3, 12)$. En efecto, la distribución final sería entonces Beta con parámetros $3 + 5 = 8$ y $12 + 20 - 5 = 27$.



En la representación gráfica de las distribuciones inicial, intermedia (final del Ejemplo 5.4.2, e inicial del 5.4.3) y final, puede observarse como la información proporcionada por los resultados experimentales va concentrando sucesivamente la densidad de probabilidad de θ : a medida que obtuviésemos más información, la densidad final de θ se iría modificando hasta concentrarse totalmente sobre el verdadero valor de θ .

Supongamos ahora que, con objeto de mejorar nuestra información sobre una magnitud μ , realizamos n medidas independientes x_1, x_2, \dots, x_n cada una de las cuales tiene una distribución normal centrada en μ y con desviación típica conocida σ . Si la información inicial sobre μ es descrita por la distribución inicial $p(\mu)$, la correspondiente distribución final será, en virtud de 5.2.2 y del Teorema de Bayes, de la forma

$$p(\mu|x_1, x_2, \dots, x_n) \propto p(x_1, x_2, \dots, x_n|\mu) p(\mu) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} p(\mu) \quad (12)$$

omitiendo innecesarias constantes de proporcionalidad.

Puede comprobarse además que

$$\sum (x_i - \mu)^2 = n\{s^2 + (\bar{x} - \mu)^2\}$$

donde $\bar{x} = \sum x_i/n$ y $s^2 = \sum (x_i - \bar{x})^2/n$. En consecuencia, la expresión (12) puede ponerse en la forma

$$p(\mu|x_1, x_2, \dots, x_n) \propto \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right\} p(\mu) \quad (13)$$

Si, además, la distribución inicial es de la familia normal, $p(\mu) = N(\mu|\mu_0, \sigma_0)$, resulta

$$p(\mu|x_1, \dots, x_n) \propto \exp \left\{ -\frac{1}{2} \left[\frac{n}{\sigma^2} (\bar{x} - \mu)^2 + \frac{1}{\sigma_0^2} (\mu - \mu_0)^2 \right] \right\}$$

Mediante el uso de la identidad

$$A(x - \alpha)^2 + B(x - \beta)^2 = (A + B) (x - m)^2 + \frac{AB}{A + B} (\alpha - \beta)^2$$

$$m = (A\alpha + B\beta)/(A + B) \quad (14)$$

la expresión anterior puede ser expresada en la forma

$$p(\mu|x_1, \dots, x_n) \propto \exp \left\{ -\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right\} \quad (15)$$

donde

$$(1/\sigma_n^2) = (1/\sigma_0^2) + (n/\sigma^2) \quad (16)$$

$$\mu_n = \sigma_n^2 \{ (\mu_0/\sigma_0^2) + (n\bar{x}/\sigma^2) \} \quad (17)$$

Comparando (15) con la definición de una densidad Normal y definiendo las precisiones respectivas (*) mediante $b_n = 1/\sigma_n^2$, $b_0 = 1/\sigma_0^2$ y $b = 1/\sigma^2$ obtenemos finalmente

$$p(\mu|x_1, \dots, x_n) = N(\mu|\mu_n, \sigma_n) \quad (18)$$

$$\sigma_n = 1/\sqrt{(b_n)}, \quad b_n = b_0 + nb \quad (19)$$

$$\mu_n = (b_0\mu_0 + n\bar{x}b)/(b_0 + nb) \quad (20)$$

(*) Para cualquier cantidad aleatoria la *precisión* se define como el recíproco de la varianza; cuanto más concentrada esté una distribución, mayor será su precisión. Frecuentemente, en el modelo normal, es más fácil expresar los resultados en términos de precisiones que en términos de varianzas o desviaciones típicas.

Puede observarse en (19) que la precisión final b_n es la suma de la precisión inicial b_0 y la proporcionada por los datos nb , y en (20) que el valor más probable μ_n de la distribución final es la *media ponderada* del valor más probable μ_0 de la distribución inicial y de la media muestral \bar{x} , con pesos proporcionales a las respectivas precisiones. Ambos hechos describen la forma general en que la distribución final *combina* la información inicial y la que proporciona el experimento.

Puede observarse además que, puesto que la distribución final de μ solo depende de la muestra $\{x_1, \dots, x_n\}$ a través de la media muestral \bar{x} y de su tamaño n , el par (n, \bar{x}) resume toda la información relevante; técnicamente, se dice que el par (n, \bar{x}) es, en este caso, un *estadístico* (esto es, una función de la muestra) *suficiente*.

La cantidad de información proporcionada por la muestra se obtiene, de acuerdo con (5), resolviendo la integral

$$I^{\mu}\{\varepsilon, N(\mu|\mu_0, \sigma_0)|x_1, x_2, \dots, x_n\} = \\ = \int N(\mu|\mu_n, \sigma_n) \log \frac{N(\mu|\mu_n, \sigma_n)}{N(\mu|\mu_0, \sigma_0)} d\mu$$

cuyo valor esperado resulta ser, simplemente,

$$I^{\mu}\{\varepsilon, N(\mu|\mu_0, \sigma_0)|n\} = \log \{\sigma_0/\sigma_n\} \quad (21)$$

esto es el logaritmo del número de veces que la muestra consigue reducir la amplitud de la desviación típica que mide nuestra incertidumbre sobre μ . Como podría esperarse, la expresión (21), que también puede escribirse como $\frac{1}{2} \log(1 + nb/b_0)$, es una función cóncava y creciente del tamaño muestral n , y una función creciente del cociente b/b_0 entre la precisión de las observaciones y la precisión inicial.

Ejemplo 5.4.4. Cantidad de tirosina (cont.)

La información inicial sobre la cantidad de tirosina μ contenida en la orina de un determinado paciente puede describirse mediante la distribución $N(\mu|39, 14,83)$ obtenida en el Ejemplo 5.1.2. Con objeto de mejorarla, se realizan tres medidas que resultan ser $x_1 = 40,62$, $x_2 = 41,8$, $x_3 = 40,44$ y que constituyen una muestra aleatoria de tamaño 3 de una población Normal centrada en μ y con desviación típica $\sigma = 2 \text{ mg}/24 \text{ h}$. Determinar la correspondiente distribución final, la información proporcionada por la muestra y la probabilidad final de que θ sea mayor de 42 mg/24 h.

La distribución inicial es $N(\mu|\mu_0, \sigma_0)$ con $\mu_0 = 39$, $\sigma_0 = 14,83$ y, por tanto,

$b_0 = 1/\sigma_0^2 = 0,0045$. La precisión de las observaciones es $b = 1/\sigma^2 = 0,25$, la media muestral es $\bar{x} = 40,953$ y su tamaño $n = 3$. En consecuencia, utilizando (19) y (20),

$$b_n = b_0 + nb = 0,0045 + 3(0,25) = 0,7545$$

$$\mu_n = (\mu_0 b_0 + nb\bar{x})/b_n = 40,94$$

y, puesto que $\sigma_n = 1/\sqrt{b_n} = 0,151$, la distribución final es $N(\mu|40,94, 0,151)$.

La información proporcionada por la muestra sobre el valor de μ será, utilizando (21), $\log(\sigma_0/\sigma_n) = 4,59$ esto es $4,59/\log 2 = 6,62$ bits; equivale por tanto a algo menos de la que proporcionarían las respuestas a *siete* preguntas binarias sobre el valor de μ , con respuestas equiprobables a priori.

5.5. Parámetros marginales

En la sección anterior hemos supuesto que, con objeto de mejorar la información de que se dispone sobre el valor θ del parámetro de interés, puede realizarse un experimento ε cuyos resultados x tienen una distribución de probabilidad $p(x|\theta)$ que sólo depende de θ . Frecuentemente, sin embargo, la distribución de los resultados experimentales x no solo depende del parámetro de interés θ sino también de otro parámetro ω , que denominaremos *parámetro marginal* (*), de forma que los resultados experimentales x tienen una distribución de probabilidad $p(x|\theta, \omega)$.

En esta situación, para poder realizar inferencias sobre el parámetro de interés θ , es necesario describir la información inicial de que se dispone, tanto sobre θ como sobre ω , mediante la correspondiente *distribución inicial conjunta* $p(\theta, \omega)$. En virtud del Teorema de Bayes, la correspondiente *distribución final conjunta* será

$$p(\theta, \omega|x) \propto p(x|\theta, \omega) p(\theta, \omega)$$

y la distribución final del parámetro de interés θ será la distribución marginal relevante, esto es

$$p(\theta|x) = \sum p(\theta, \omega_i|x)$$

si ω es discreta o,

$$p(\theta|x) = \int p(\theta, \omega|x) d\omega$$

si ω es una cantidad aleatoria continua.

(*) Algunos autores emplean el término *parámetro perturbador*.

Ejemplo 5.5.1. Diagnosis

Con objeto de determinar la causa θ de un determinado síndrome ($\theta = \theta_1, \theta = \theta_2$ o $\theta = \theta_3$) se realiza una exploración cuyo resultado puede ser positivo ($x = 1$) o negativo ($x = 0$). La distribución de probabilidad del resultado experimental depende de la causa del síndrome θ y de la posible existencia ω de una infección ($\omega = 1$ existe infección, $\omega = 0$ no existe) de manera que $p(x|\theta, \omega)$ es de la forma

$$\begin{aligned} p(x = 1|1, 1) &= 0,9 \\ p(x = 1|1, 0) &= 0,7 \\ p(x = 1|2, 1) &= 0,5 \\ p(x = 1|2, 0) &= 0,2 \\ p(x = 1|3, 1) &= 0,2 \\ p(x = 1|3, 0) &= 0,0 \end{aligned}$$

La información inicial sobre los pares (θ, ω) viene dada por la tabla

	$\theta = \theta_1$	$\theta = \theta_2$	$\theta = \theta_3$
$\omega = 1$	0,1	0,2	0,3
$\omega = 0$	0,2	0,1	0,1

Determinar la distribución final de θ según sea el resultado de la exploración.

Si el resultado es positivo ($x = 1$) tenemos, en virtud del Teorema de Bayes

$$p(\theta, \omega|x = 1) \propto p(x = 1|\theta, \omega) p(\theta, \omega)$$

Realizando operaciones y normalizando, la correspondiente distribución final conjunta viene dada por la tabla

	$\theta = \theta_1$	$\theta = \theta_2$	$\theta = \theta_3$
$\omega = 1$	0,22	0,24	0,15
$\omega = 0$	0,34	0,05	0,00

Y por lo tanto la distribución final de θ si $x = 1$ resulta ser

$$p(\theta|x = 1) = (0,56, 0,29, 0,15)$$

como $p(x = 0|\theta, \omega) = 1 - p(x = 1|\theta, \omega)$, la distribución final de θ si $x = 0$ puede determinarse de forma análoga y resulta ser

$$p(\theta|x = 0) = (0,12, 0,31, 0,57)$$

La distribución inicial de θ era $p(\theta) = (0,3, 0,3, 0,4)$ que daba a las tres posibles causas del síndrome una probabilidad parecida. Puede observarse que si el resultado de la exploración es positivo, aumenta la verosimilitud de $\theta = \theta_1$, como causa del síndrome, mientras que si el resultado es negativo aumenta la verosimilitud de $\theta = \theta_3$, como cabía esperar de la simple observación de la función de verosimilitud.

Concluremos esta sección estudiando las conclusiones que pueden obtenerse a partir de una muestra aleatoria de una población normal sobre el valor de su media μ cuando también se desconoce la desviación típica σ .

Según vimos (Ecuación 5.2.2), la función de verosimilitud correspondiente a una muestra aleatoria $z = (x_1, x_2, \dots, x_n)$ de una población normal $N(x|\mu, \sigma)$ es

$$\begin{aligned} l_z(\mu, \sigma) &= p(x_1, \dots, x_n|\mu, \sigma) = \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} \end{aligned} \quad (1)$$

Si $p(\mu, \sigma)$ describe la información inicial de que se dispone sobre los valores de μ y de σ , la distribución final de μ y σ será

$$p(\mu, \sigma|x_1, \dots, x_n) \propto p(x_1, \dots, x_n|\mu, \sigma) p(\mu, \sigma) \quad (2)$$

y la distribución final de μ

$$p(\mu|x_1, \dots, x_n) = \int p(\mu, \sigma|x_1, \dots, x_n) d\sigma \quad (3)$$

Como veremos en el capítulo próximo, la situación en que no se dispone de información inicial sobre μ y σ puede describirse, *en este contexto*, mediante la función $p(\mu, \sigma) = \sigma^{-1}$ de forma que sustituyendo en (2) y realizando la integral (3) se obtiene

$$p(\mu|x_1, x_2, \dots, x_n) \propto [1 + \{(\bar{x} - \mu)/s\}^2]^{-n/2} \quad (4)$$

$$\bar{x} = \sum x_i/n, \quad s^2 = \sum (x_i - \bar{x})^2/n \quad (5)$$

Comparando (4) con la Definición 4.3.6 de una distribución Student resulta que, en ausencia de información inicial

$$p(\mu|x_1, x_2, \dots, x_n) = St(\mu|\bar{x}, s/\sqrt{(n-1)}, n-1) \quad (6)$$

Ejemplo 5.5.2. pH de la saliva

Para la elaboración de un determinado preparado farmacológico resulta necesario precisar el grado de acidez (valor pH) medio de la saliva de los pacientes a que va dirigido. Tomada una muestra aleatoria de 10 pacientes se obtienen los valores

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
5,60	5,95	6,01	6,06	6,52	5,85	6,15	6,43	5,34	6,41

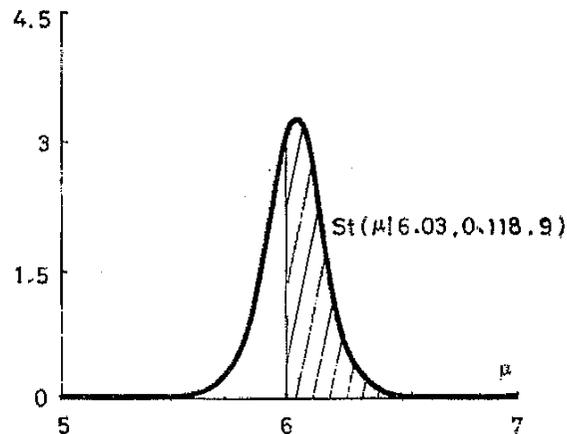
Suponiendo que los datos proceden de una distribución normal $N(x|\mu, \sigma)$, y en ausencia de información inicial sobre sus parámetros, determinar la distribución final de μ y la probabilidad de que $\mu > 6$.

Utilizando las expresiones (5), se obtiene en este caso

$$\bar{x} = 6,03, \quad s^* = 0,126$$

y por tanto $s/\sqrt{(n-1)} = s/3 = 0,118$ de forma que la distribución pedida es

$$p(\mu|x_1, \dots, x_{10}) = St(\mu|6,03, 0,118, 9)$$



Finalmente,

$$p[6 < \mu < \infty | x_1, \dots, x_n] = \Phi_9(\infty) - \Phi_9\left(\frac{6 - 6,03}{0,118}\right) = 1 - \Phi_9(-0,25) = \Phi_9(0,25)$$

donde Φ_9 es la función de distribución de la t de Student normalizada con 9 grados de libertad. Haciendo uso de las tablas correspondientes se obtiene $\Phi_9(0,25) \approx 0,804$, que es la probabilidad pedida.

5.6. Predicción

Frecuentemente, el parámetro de interés en un problema de decisión no es el parámetro de una distribución sino el resultado de una observación futura. Así por ejemplo, el suceso incierto relevante para decidir si someter o no a un nuevo paciente a una determinada operación es el estado final ($x = 1$ satisfactorio, $x = 0$ insatisfactorio) de ese paciente y no la proporción θ de pacientes con los que se obtiene un resultado satisfactorio. Sin embargo, la información acumulada sobre el valor de θ mediante el Teorema de Bayes puede ser utilizada para *predecir* el posible resultado de operar al nuevo paciente.

En general, si x_1, x_2, \dots, x_n son resultados independientes con una distribución común $p(x_i|\theta)$, y $p(\theta)$ describe la información inicial sobre el valor de θ , la información acumulada sobre θ vendrá dada, en virtud del Teorema de Bayes por

$$p(\theta|x_1, x_2, \dots, x_n) \propto \prod_{i=1}^n p(x_i|\theta) p(\theta)$$

Esta información puede combinarse con la distribución $p(x|\theta)$ de una nueva observación x para describir la información que se posee sobre sus posibles valores. En efecto,

$$p(x|x_1, x_2, \dots, x_n) = \sum p(x|\theta_i) p(\theta_i|x_1, x_2, \dots, x_n) \quad (1)$$

si θ es una cantidad aleatoria discreta y

$$p(x|x_1, x_2, \dots, x_n) = \int p(x|\theta) p(\theta|x_1, x_2, \dots, x_n) d\theta \quad (2)$$

si es continua, proporciona una distribución de x totalmente conocida, con la que podemos hacer predicciones. Esta distribución recibe el nombre de *distribución predictiva final*. Su semejanza con la distribución predictiva (inicial) estudiada en la Sección 5.3 es obvia; tan solo hemos sustituido en las expresiones 5.3.1 y 5.3.2 la distribución inicial $p(\theta)$ por la distribución final $p(\theta|x_1, x_2, \dots, x_n)$. Con la distribución predictiva inicial $p(x)$ pueden hacerse predicciones sobre el resultado experimental x antes de realizar observación alguna, y puede utilizarse para analizar las implicaciones de la distribución inicial. En la distribución predictiva final $p(x|x_1, x_2, \dots, x_n)$ se recoge la información proporcionada por las observaciones x_1, x_2, \dots, x_n y se utiliza para predecir el resultado de la próxima observación x . Se trata aquí de un problema de inferencia que solo puede ser resuelto una vez analizada la obtención de la distribución final.

Ejemplo 5.6.1. *Diagnosis (cont.)*

Determinar la distribución predictiva final de los resultados de los tests descritos en el Ejemplo 5.2.1 utilizando la distribución final sobre las causas de las dolencias obtenidas en el Ejemplo 5.4.1 cuando ambos tests resultan positivos.

Si el resultado del test es (1, 1) la distribución final de θ resultaba ser

$$p(\theta|1, 1) = \{0.459, 0.199, 0.077, 0.199, 0.066\}$$

y, en consecuencia, utilizando esta distribución en lugar de la distribución inicial $p(\theta)$ y procediendo como en el Ejemplo 5.3.1 se obtiene

$$p(1, 1|1, 1) = 0.345$$

$$p(1, 0|1, 1) = 0.181$$

$$p(0, 1|1, 1) = 0.374$$

$$p(0, 0|1, 1) = 0.100$$

Comparando este resultado con la distribución predictiva inicial obtenida en el Ejemplo 5.3.1

$$p(1, 1) = 0.181$$

$$p(1, 0) = 0.235$$

$$p(0, 1) = 0.484$$

$$p(0, 0) = 0.100$$

puede observarse que, como era de esperar, el hecho de que (1, 1) haya sucedido una vez hace *más* probable que vuelva a suceder de nuevo. Esta probabilidad solo se man-

tendría constante si el valor de θ fuese conocido, en cuyo caso $p(1, 1)$ sería constantemente igual a $p(1, 1|\theta)$ puesto que nuevos resultados experimentales no podrían ya mejorar nuestra información sobre θ .

Concluiremos esta sección considerando distribuciones predictivas finales correspondientes a las dos situaciones experimentales repetidamente mencionadas en este capítulo, descritas respectivamente por las distribuciones Binomial y Normal.

Considérese que se ha obtenido una muestra aleatoria de tamaño n de una población y que se ha encontrado que r elementos de la muestra poseían una determinada característica. Se desea saber la probabilidad de que un nuevo elemento de la población escogido al azar tenga ($x = 1$) o no tenga ($x = 0$) la característica mencionada. Si θ es la proporción (desconocida) de elementos de la población que tienen tal característica tenemos

$$p(x = 1|\theta) = \theta$$

$$p(x = 0|\theta) = 1 - \theta$$

En consecuencia, si $p(\theta|r, n)$ es la distribución final que describe la información de que se dispone sobre θ tras observar los resultados experimentales, tendremos en virtud de (2)

$$p(x = 1|r, n) = \int \theta p(\theta|r, n) d\theta = E(\theta|r, n)$$

y, naturalmente, $p(x = 0|r, n) = 1 - p(x = 1|r, n)$, de forma que la probabilidad de que un nuevo elemento de la población tenga la característica estudiada es simplemente la *media* de la distribución final de θ .

En particular, si la distribución inicial de θ es $Be(\theta|\alpha, \beta)$ y por tanto (Ecuación 5.4.6) su distribución final es $Be(\theta|\alpha + r, \beta + n - r)$, la probabilidad buscada es, en virtud del Teorema 4.5.3,

$$p(x = 1|r, n) = \frac{\alpha + r}{\alpha + \beta + n} \quad (3)$$

que, para valores grandes de r y de n , es aproximadamente igual a la proporción muestral r/n .

Ejemplo 5.6.2. *Tasa de morbilidad (cont.)*

Determinar la probabilidad de que un elemento de la población escogido al azar esté afectado de gripe si la distribución inicial sobre la proporción θ

de elementos en la población con grupe es la distribución $Be(\theta|3, 12)$ construida en el Ejemplo 5.1.4 y una muestra de 20 elementos contenía 5 enfermos.

Utilizando la ecuación (3)

$$p(x = 1|5, 20) = \frac{3 + 5}{3 + 12 + 20} = \frac{8}{35} = 0,23$$

que es la media de la distribución final $Be(\theta|8, 27)$ descrita en el Ejemplo 5.4.3.

Considérese finalmente que se ha obtenido una muestra aleatoria (x_1, x_2, \dots, x_n) de una población Normal de media μ y desviación típica conocida σ ; se desea determinar la distribución predictiva final de una nueva observación x . Por hipótesis

$$p(x|\mu) = N(x|\mu, \sigma)$$

En consecuencia, si $p(\mu|x_1, \dots, x_n)$ es la distribución final que describe la información de que se dispone sobre μ tras observar los resultados experimentales, tendremos, en virtud de (2),

$$p(x|x_1, \dots, x_n) = \int N(x|\mu, \sigma) p(\mu|x_1, \dots, x_n) d\mu$$

En particular, si la distribución inicial de μ es $N(\mu|\mu_0, \sigma_0)$ y por tanto (Ecuación 5.4.18) la distribución final es $N(\mu|\mu_n, \sigma_n)$ donde μ_n y σ_n vienen dados por las ecuaciones 5.4.19 y 5.4.20, resulta

$$p(x|x_1, \dots, x_n) = \int N(x|\mu, \sigma) N(\mu|\mu_n, \sigma_n) d\mu = N(x|\mu_n, \sqrt{\sigma^2 + \sigma_n^2}) \quad (4)$$

de forma análoga a la Ecuación 5.3.8. Comparando la ecuación (4) con la 5.3.8 se observa que, como era de esperar, la distribución predictiva final es de la misma forma que la distribución predictiva inicial con los parámetros de la distribución final de μ en lugar de los de su distribución inicial.

Comparando (4) con la distribución final de μ , $N(\mu|\mu_n, \sigma_n)$, se observa que tiene la misma media y una desviación típica necesariamente mayor. Cuando el tamaño muestral n crece, la desviación típica de la distribución final σ_n tiende a cero, lo que nos permite conocer el valor de μ con una precisión arbitrariamente grande; sin embargo, cuando n crece la desviación típica de la distribución predictiva final solo tiende a σ por lo que no podemos aspirar a predecir con mayor precisión el resultado de una observación futura.

Ejemplo 5.6.3. Cantidad de tirosina (cont.)

Determinar la distribución predictiva de la próxima medición de la cantidad de tirosina μ en la orina de un paciente si tales mediciones se suponen normalmente distribuidas con media μ y desviación típica $\sigma = 2$, la distribución inicial de μ es la $N(\mu|39, 14,83)$ obtenida en el Ejemplo 5.1.2 y se llevan realizadas tres mediciones, la media aritmética de cuyos resultados es 40,953. Determinar la probabilidad de que tal medición resulte superior a 42 mg/24 h.

La correspondiente distribución final, obtenida en el Ejemplo 5.4.4, es $p(\mu|x_1, x_2, x_3) = N(\mu|40,49, 1,151)$; en consecuencia, en virtud de (4), la distribución pedida será

$$p(x|x_1, x_2, x_3) = N(x|40,94, \sqrt{2^2 + 1,151^2}) = N(x|40,94, 2,308)$$

Finalmente,

$$\begin{aligned} p(x > 42|x_1, x_2, x_3) &= p[42 < x < \infty|x_1, x_2, x_3] = \\ &= \Phi(\infty) - \Phi\left(\frac{42 - 40,94}{2,308}\right) = 1 - \Phi(0,46) = 0,3230 \end{aligned}$$

5.7. Discusión y referencias

Tradicionalmente, el problema de inferencia se formulaba como el problema de extraer conclusiones probabilísticas sobre los valores de los parámetros de una distribución basándose en la observación de una muestra extraída de ella. En este Capítulo hemos desarrollado la solución *Bayesiana* a este problema que, como hemos visto, consiste en acumular a la información inicial la información proporcionada por los datos mediante el uso del Teorema de Bayes. Esta solución es la única compatible con los principios de coherencia expuestos en el Capítulo 2.

Existe una bibliografía creciente sobre los métodos estadísticos Bayesianos. Resultan ya clásicos los textos de Jeffreys (1939/1967), Good (1950, 1965), Lindley (1965), De Finetti (1970/1975, 1972), DeGroot (1970), Zellner (1971), y Box & Tiao (1973). Entre los libros de texto más elementales recientemente aparecidos podemos citar los de Novick & Jackson (1974), Phillips (1973), Savage (1968), Schmitt (1969) y Winkler (1972). Sin embargo, una proporción importante de los resultados conocidos en inferencia Bayesiana hay que buscarlos todavía en las revistas especializadas; Lindley (1971 a) proporciona un detallado análisis de los publicados antes de esa fecha.

Recientemente, han aparecido varias colecciones de artículos que recogen aportaciones importantes a la metodología Bayesiana. Merecen especial aten-

ción los editados por Godambe & Sprott (1971), Aykac & Brumat (1977), Fienberg & Zellner (1975), Harper & Hooker (1976), Barra *et. al.* (1977), Zellner (1980) y Bernardo *et. al.* (1980).

La escuela clásica de inferencia se basa en principios radicalmente diferentes; no se deriva de conjunto axiomático alguno y viola, en particular, los principios de coherencia mencionados en la Sección 2.3. Existen varios textos que permiten un análisis comparativo de ambas metodologías; podemos citar entre ellos los de Barnett (1973), Cox & Hinklev (1974) y DeGroot (1975) que adoptan, respectivamente, posturas ecléctica, clásica y bayesiana.

PROBLEMAS

1. Se sabe que la cantidad de gamma glutamil-transpeptidasa (γ GT) en el suero sanguíneo en varones, θ , tiene un valor medio de 22.75 mU/ml y que, con probabilidad 0.99, es menor que 42.75. Determinar la distribución normal que mejor describe esta información.
2. Se sabe que la proporción θ de personas cuyo grupo sanguíneo es el A está entre el 30 y el 70 % con probabilidad 0.90. Determinar, utilizando tablas, una distribución Beta que describa adecuadamente esta información. Comparar este resultado con el que se obtiene utilizando la transformación logarítmica para normalizar.
3. En una investigación sobre la proporción θ de personas afectadas por una epidemia se realiza un muestreo de 100 personas elegidas al azar, 8 de las cuales resultan estar afectadas. Construir y representar la correspondiente función de verosimilitud.
4. Construir y representar la función de verosimilitud correspondiente a las medidas $x_1 = 208$, $x_2 = 215$, $x_3 = 198$, $x_4 = 218$ mg/ml obtenidas de una población $N(\mu, 10)$ al investigar el colesterol contenido en el suero sanguíneo de un paciente.
5. Para determinar la cantidad de enzima γ GT, θ , contenida en el suero sanguíneo de un determinado paciente va a realizarse una medida con un aparato que produce medidas normales, centradas en θ con desviación típica igual a 3 mU/ml. Determinar la correspondiente distribución predictiva utilizando la distribución inicial para θ del Problema 1.
6. La información inicial sobre el colesterol μ contenido en la sangre de un paciente puede describirse mediante una distribución normal $N(\mu|200, 16)$. Determinar la correspondiente distribución final una vez observados los resultados descritos en el Problema 4.
7. Se sabe que la proporción θ de personas afectadas por una epidemia puede describirse mediante una distribución Beta de media 0.1 y desviación típica 0.05. Determinar la distribución final de θ una vez observados los resultados descritos en el Problema 3.

8. Al medir la cantidad de glicerol libre μ en la sangre de un paciente se obtienen los valores 1.1, 1.0, 1.3, 0.9 y 1.2 mg/100 ml. Determinar la distribución final de μ en ausencia de información tanto sobre μ como sobre la precisión de los datos.
9. Se sabe que las proteínas contenidas en el líquido cefalorraquídeo se sitúan entre 15 y 40 mg/100 ml con probabilidad 0.99. Realizadas dos punciones a un determinado paciente y analizados los resultados se obtienen valores de 26 y 23 mg/100 ml con un método que proporciona resultados con distribución normal centrados en el verdadero valor y con desviación típica igual a 3 mg/100 ml. Determinar la distribución predictiva sobre el resultado de una tercera punción.
10. La efectividad de dos tratamientos alternativos, μ_1 y μ_2 , para una determinada enfermedad cerebral depende del contenido θ de ClNa en el líquido cefalorraquídeo (LCR). Expresándola en años esperados de vida, la utilidad de cada uno de ellos viene dada por

$$u(t_1, \theta) = 12 - (\theta - 720)/10$$

$$u(t_2, \theta) = 12 - (750 - \theta)/8$$

Se sabe que el valor de θ se sitúa entre 720 y 750 mg/100 ml con probabilidad 0.999. Realizados dos análisis del LCR del paciente mediante un método con desviación típica 2 mg/100 ml se obtienen los valores 741 y 743 mg/100 ml. Determinar el tratamiento óptimo.

Métodos aproximados de inferencia

En este capítulo se describen un conjunto de métodos que permiten obtener *conclusiones aproximadas* sobre el valor de la magnitud de interés, derivadas de su distribución final o de aproximaciones a ella.

Como *medidas descriptivas* de la distribución final se utilizan algunas de sus *características* (media, moda, desviación típica...) y las llamadas *regiones creíbles* o de confianza.

Se describen algunos métodos para transformar la distribución final en otra *aproximadamente normal*, con objeto de facilitar su análisis. Se estudia el *comportamiento asintótico* de la distribución final a medida que crece el tamaño de la muestra.

Se define la *distribución final de referencia* que describe las conclusiones que pueden extraerse de los datos cuando no se dispone de información inicial.

Se comenta finalmente la *estabilidad* de la distribución final frente a cambios en los datos, el modelo, o la distribución inicial.

El cálculo exacto de la distribución final de la magnitud de interés es frecuentemente muy complicado; además, incluso en aquellos casos en que puede obtenerse, la distribución final resulta a menudo un concepto demasiado complejo para ser utilizado por quienes no poseen un cierto grado de sofisticación matemática. Resulta en consecuencia deseable disponer de métodos que permitan obtener *conclusiones aproximadas* sobre el valor de la magnitud de interés fáciles de obtener y sencillas de interpretar. Una forma de hacerlo es definir características fácilmente interpretables de la distribución final, que describan aproximadamente su localización y su forma.

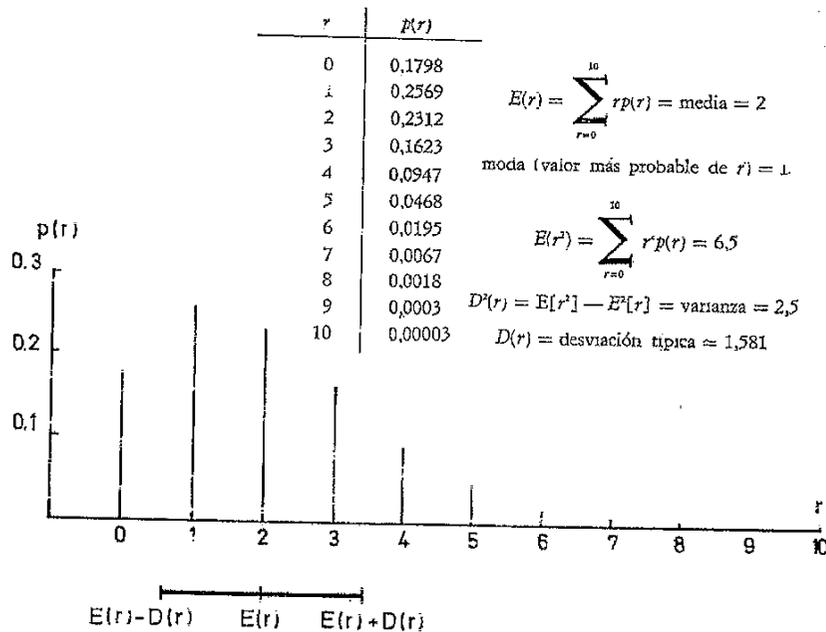
6.1. Descripción de la distribución final

Según vimos en el Capítulo 2, la *única* forma coherente de expresar la información de que se dispone sobre el valor de una magnitud incierta θ consiste en la especificación de su distribución de probabilidad.

Cuando la magnitud de interés θ es *discreta*, de forma que el conjunto de valores posibles de θ es finito o infinito numerable, la forma más sencilla de describir su distribución es mediante la correspondiente función de probabilidad $p(\theta_i) = p(\theta = \theta_i)$. La comprensión de su significado y alcance puede facilitarse mediante su representación gráfica y el cálculo de algunas de sus características (ver Sección 4.5) como su media, moda o desviación típica.

Ejemplo 6.1.1. Tasas de morbilidad (cont.)

Representar gráficamente y determinar las características de la distribución $Bb(r|3, 12, 10)$, obtenida en el Ejemplo 5.3.2, que describe nuestra información sobre el número de personas afectadas por la epidemia en una muestra aleatoria de tamaño 10.



La simple observación de la representación gráfica, donde además de la función de probabilidad se ha dibujado una desviación típica a cada lado de la media, proporciona una clara idea intuitiva de la información de que se dispone sobre el valor de r. Los momentos, pueden también ser calculados directamente a partir de las propiedades de la distribución Beta-binomial, en efecto si $p(x) = Bb(x|a, b, n)$, entonces (Raiffa & Schlaifer, 1961, p. 237)

$$E(x) = \frac{na}{a+b}$$

$$D^2(x) = \frac{nab}{(a+b)^2} \left(\frac{n+a+b}{a+b+1} \right)$$

Sustituyendo por $a = 3, b = 12$ y $n = 10$ se reproducen los valores encontrados.

Cuando el parámetro de interés es una magnitud continua, la forma más sencilla de describir su distribución es mediante la correspondiente densidad de probabilidad. La asimilación intuitiva de su contenido se facilita enormemente mediante su representación gráfica, el cálculo de sus principales características, y la determinación de un cierto número de *regiones creíbles*.

DEFINICIÓN 6.1.1. Sea $\theta \in \Theta$ una cantidad aleatoria continua con densidad de probabilidad $p(\theta)$. La región creíble de nivel de confianza p, que representaremos con $I_p(\theta)$, es aquel subconjunto del espacio paramétrico Θ de longitud mínima entre todos aquellos cuya probabilidad asociada es p.

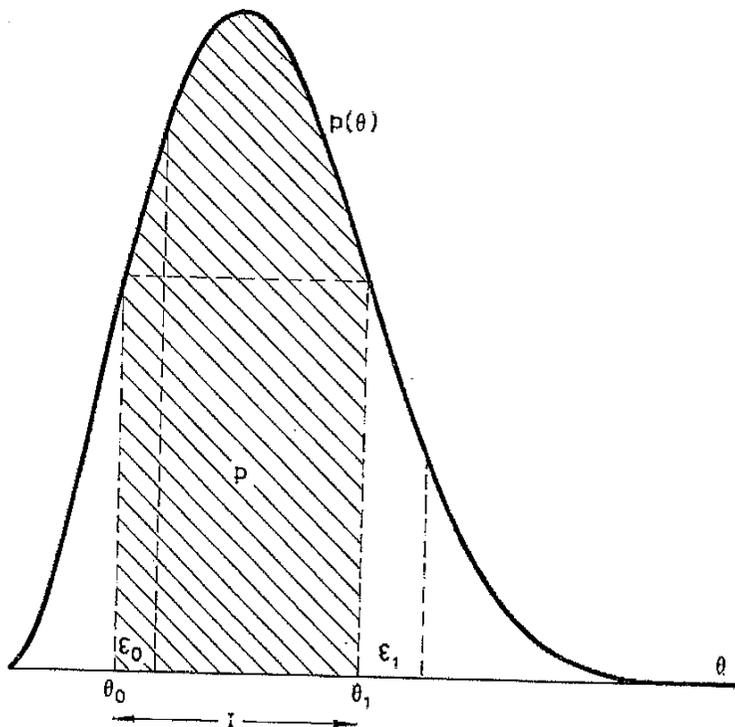
Intuitivamente, una región creíble es la que contiene a aquellos valores de θ que resultan más *probables* a la vista de la información de que se dispone. En efecto,

TEOREMA 6.1.1. Cualquiera que sea p, todos los puntos que pertenecen a la región creíble $I_p(\theta)$ tienen mayor densidad de probabilidad que cualquiera de los puntos que no pertenecen a $I_p(\theta)$.

Demostración

Consideremos la región I obtenida al seccionar la densidad de probabilidad $p(\theta)$ mediante una paralela al eje de abscisas, y sea p la probabilidad asociada a I; en la figura, la región I está constituida por el intervalo (θ_0, θ_1) mientras que p es el área de la región sombreada.

Si modificamos la región I manteniendo constante la probabilidad asociada a ella (por ejemplo desplazando a la derecha el intervalo i de la figura), aumentamos necesariamente su longitud. En efecto, para cualquier desplazamiento $\epsilon_1 > 0$ del punto θ_1 , el área incorporada será, haciendo uso del teorema del valor medio, de la forma $\epsilon_1 p(\theta_1^*)$ con $\theta_1^* \in (\theta_1, \theta_1 + \epsilon_1)$ y el área eliminada de la forma $\epsilon_0 p(\theta_0^*)$ con $\theta_0^* \in (\theta_0, \theta_0 + \epsilon_0)$; puesto que ambas áreas deben ser iguales y si, como en la figura, suponemos $p(\theta_1^*) < p(\theta_0^*)$



resulta $\epsilon_1 > \epsilon_0$ y por tanto la longitud del nuevo intervalo es mayor que la del original. Desplazando el intervalo hacia la izquierda se obtendrá un resultado análogo.

En consecuencia, la región I es la región de longitud mínima entre las que tienen la misma probabilidad asociada y se trata por tanto de la región creíble de nivel de confianza p . Recíprocamente, una región creíble debe ser de la forma descrita en el teorema porque, de otra manera, un pequeño desplazamiento disminuiría su longitud, en contra de la definición de región creíble.

Además de una interpretación intuitiva del significado de las regiones creíbles, el teorema anterior proporciona un método gráfico de construir las. En efecto, para construir una región creíble de nivel p basta trazar paralelas al eje de abscisas, progresivamente más bajas, hasta que la probabilidad asociada a la región que determinan sea precisamente el nivel pedido p . Por otra parte, el método de construcción mencionado permite apreciar fácilmente la

unicidad de la solución, esto es el hecho de que para cada nivel solo existe una región creíble.

Frecuentemente, las distribuciones finales que aparecen en la práctica son unimodales. En este caso, el Teorema 6.1.1 nos permite asegurar que las regiones creíbles son intervalos. A estas regiones creíbles conexas se les llama también *intervalos de confianza*.

Si además de unimodal la distribución final es simétrica; todos los intervalos de confianza estarán obviamente centrados en el valor medio de la distribución, esto es, serán de la forma $[E(\theta) - r, E(\theta) + r]$ donde el *radio* r del intervalo dependerá del nivel elegido. De acuerdo con la Definición 6.1.1. el intervalo de nivel p debe contener probabilidad p y por tanto

$$P[E(\theta) - r < \theta < E(\theta) + r] = p$$

o, alternativamente, usando la simetría alrededor de $E(\theta)$

$$P[\theta < E(\theta) + r] = (1 + p)/2$$

Si $\omega = \{\theta - E(\theta)\}/D(\theta)$ es la correspondiente magnitud *tipificada*, tendremos $p\{\omega < r/D(\theta)\} = (1 + p)/2$ y por lo tanto

$$F_{\omega}\{r/D(\theta)\} = (1 + p)/2 \tag{1}$$

donde F_{ω} es la función de distribución de ω . Resolviendo en r esa ecuación por medio de las tablas correspondientes puede construirse el intervalo.

En particular, si la distribución final es Normal, tendremos $p(\omega) = N(\omega|0, 1)$ y la ecuación (1) será $\Phi\{r/D(\theta)\} = (1 + p)/2$. Como puede comprobarse en las Tablas correspondientes, los valores n_p para los que se verifica la ecuación $\Phi(n_p) = (1 + p)/2$ son, para distintos valores de p ,

p	0.5	0.75	0.90	0.95	0.99	0.999
n_p	0,6745	1,1503	1,6449	1,9600	2,5758	3,2905

de manera que $r = n_p D(\theta)$ y los correspondientes intervalos de confianza serán de la forma

$$I_p(\theta) = [E(\theta) - n_p D(\theta), E(\theta) + n_p D(\theta)] \tag{3}$$

En general, las regiones de confianza correspondientes a distribuciones simétricas serán de la forma

$$I_p(\theta) = [E(\theta) - x_p D(\theta), E(\theta) + x_p D(\theta)] \quad (4)$$

donde x_p es la solución de la ecuación

$$F_{\omega}(x_p) = (1 + p)/2 \quad (5)$$

y F_{ω} es la función de distribución de la variable tipificada $\omega = \{\theta - E(\theta)\}/D(\theta)$. Generalmente la solución de la ecuación (5) se encuentra utilizando las tablas estadísticas relevantes (*).

Para que los resultados finales de una investigación estadística resulten fácilmente asimilables a nivel intuitivo, la descripción analítica de la distribución final correspondiente debe complementarse con la representación gráfica de su densidad de probabilidad; es conveniente además representar sobre la misma figura los valores de las características principales de la distribución así como los de un conjunto de regiones creíbles de niveles convenientemente escogidos. Es frecuente elegir a estos efectos los niveles 0,5, 0,9, 0,95 y 0,99.

Ejemplo 6.1.2. Cantidad de tirosina (cont.)

Determinar las regiones creíbles de niveles 0,5, 0,9, 0,95 y 0,99 correspondientes a la distribución final $N(\mu|40,94, 0,151)$ sobre la cantidad de tirosina de un determinado paciente encontrada en el Ejemplo 5.4.4. Representar gráficamente tales regiones con relación a la densidad de probabilidad que las origina.

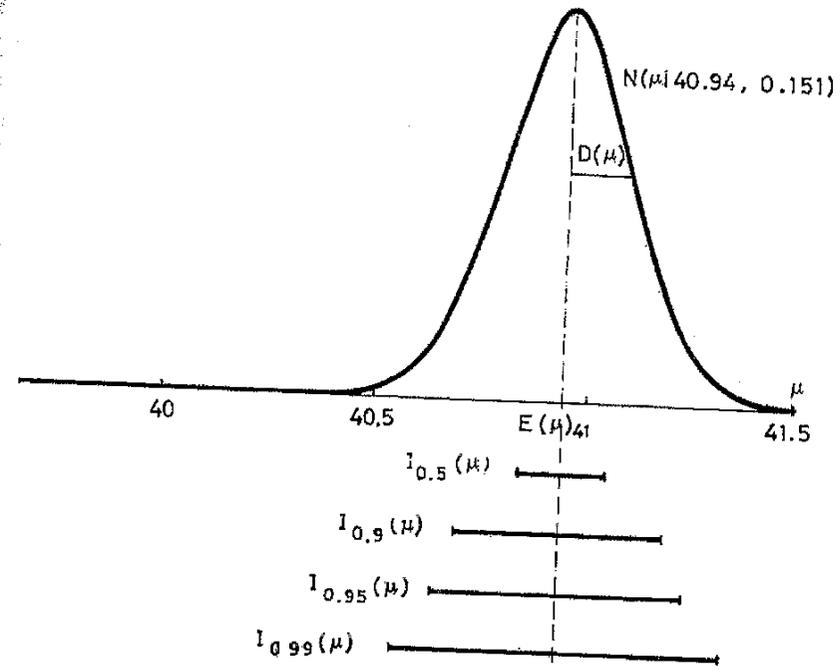
Puesto que la distribución Normal es unimodal y simétrica, tanto la *moda* como la *mediana* coinciden con la *media* μ y las regiones creíbles están constituidas por *intervalos* centrados en ese valor.

Utilizando la Ecuación (3) y la Tabla (2), las regiones creíbles son los intervalos

p	0,5	0,9	0,95	0,99
$I_p(\mu)$	(40,84, 41,04)	(40,69, 41,19)	(40,64, 41,24)	(40,55, 41,33)

La representación gráfica correspondiente, que resulta ser,

(*) La definición de región creíble ser extendida al caso discreto; sin embargo, cuando la cantidad aleatoria es discreta no existe región creíble para cualquier nivel de confianza, sino sólo para algunos de ellos, debido a las discontinuidades de la función de distribución.



permite una asimilación rápida e intuitiva de la información sobre los posibles valores de μ contenida en su distribución final.

6.2. Familias conjugadas de distribuciones

En el capítulo anterior, se han mencionado varios ejemplos en los que la elección de una determinada familia para representar la información inicial conducía a unos resultados matemáticos especialmente sencillos. Así, en la Sección 5.4 encontrábamos que si la información inicial sobre el parámetro θ de una distribución binomial se describía mediante una distribución de la familia Beta, la correspondiente distribución final pertenecía a esta misma familia, y toda la información proporcionada por los datos podía resumirse en el par (r, n) ; análogamente, si la información inicial sobre la media μ de una distribución normal con desviación típica conocida se podía describir mediante una distribución normal, entonces la correspondiente distribución final resultaba ser asimismo normal y toda la información proporcionada por la muestra podía resumirse en el par (\bar{x}, n) . Estos ejemplos son casos particu-

lares de un amplio espectro de situaciones en las que la *aproximación*, mediante distribuciones pertenecientes a una determinada familia, de la distribución que describe la información inicial, permite obtener resultados sencillos, de tipo analítico, en problemas que de otra manera requerirían técnicas de integración numérica y el uso sistemático de un ordenador.

DEFINICIÓN 6.2.1. Se dice que una familia de distribuciones de θ es conjugada con respecto a un determinado modelo probabilístico $p(x|\theta)$ si para cualquier distribución inicial perteneciente a tal familia se obtiene una distribución final que también pertenece a ella.

Como vimos en la Sección 5.4, la familia de distribuciones Beta es conjugada con respecto al modelo probabilístico de Bernoulli (y por lo tanto con respecto a todos los modelos probabilísticos proporcionales a él, como el Binomial), y la familia de distribuciones normales es conjugada con respecto al modelo probabilístico normal con desviación típica conocida. Más adelante, en el Teorema 6.2.2, se citan otros ejemplos.

DEFINICIÓN 6.2.2. Se dice que un estadístico (esto es, una función de la muestra), $t = t(x_1, \dots, x_n)$, es suficiente para hacer inferencias sobre los parámetros $(\theta_1, \dots, \theta_k)$ de un modelo probabilístico $p(x|\theta_1, \dots, \theta_k)$ si cualquiera que sea la distribución inicial $p(\theta_1, \dots, \theta_k)$ de estos parámetros, su distribución final sólo depende de dicho estadístico, esto es $p(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = p(\theta_1, \dots, \theta_k|t)$.

El par $(\sum x_i, n)$ es suficiente si el modelo es de Bernoulli. En efecto, si observamos una muestra aleatoria (x_1, \dots, x_n) de forma que para todo x_i ,

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0, 1$$

y la distribución inicial de θ es $p(\theta)$, la correspondiente distribución final será

$$\begin{aligned} p(\theta|x_1, \dots, x_n) &\propto p(\theta) \Pi \theta^{x_i}(1-\theta)^{1-x_i} \\ &= p(\theta) \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \end{aligned}$$

y por lo tanto, con $r = \sum x_i$ y cualquiera que sea $p(\theta)$, la distribución final de θ , $p(\theta|x_1, \dots, x_n)$ solo depende de la muestra (x_1, \dots, x_n) a través del par (r, n) que es, por tanto, un estadístico suficiente.

Análogamente, el par (\bar{x}, n) es un estadístico suficiente para el modelo normal con desviación típica conocida. En efecto, si (x_1, \dots, x_n) es una muestra aleatoria de una población $N(x|\mu, \sigma)$ con σ conocido y $p(\mu)$ es la distribución inicial de μ , la distribución final de μ resulta ser (Ec. 5.4.13)

$$p(\mu|x_1, \dots, x_n) \propto p(\mu) \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right\}$$

que únicamente depende de la muestra (x_1, \dots, x_n) a través del par (\bar{x}, n) .

La existencia de un estadístico suficiente está íntimamente ligada a la forma de la función de verosimilitud. En efecto,

TEOREMA 6.2.1. Bajo condiciones muy generales (*), el par $\{t(x_1, \dots, x_n), n\}$ es un estadístico suficiente para hacer inferencias sobre el parámetro θ de un modelo $p(x|\theta)$ basadas en la muestra $z = \{x_1, \dots, x_n\}$ si, y solamente si, la función de verosimilitud correspondiente, $p(z|\theta)$ es de la forma

$$p(z|\theta) = F(z)G(\theta)^n \exp\{t(z)\zeta(\theta)\} \quad (1)$$

donde F, G y ζ son funciones arbitrarias.

La demostración excede ampliamente los límites matemáticos elegidos para este libro (**); es fácil comprobar sin embargo que los modelos mencionados cumplen efectivamente la condición del Teorema. Así, por ejemplo, en el caso Bernoulli tenemos

$$p(z|\theta) = \Pi \theta^{x_i}(1-\theta)^{1-x_i} = (1-\theta)^n \exp\{(\sum x_i) \log\{\theta/(1-\theta)\}\}$$

que es de la forma (1) con $F(z) = 1$, $G(\theta) = (1-\theta)$, $t(z) = \sum x_i$ y $\zeta(\theta) = \log\{\theta/(1-\theta)\}$.

Los modelos que pueden ser expresados en la forma (1) constituyen la llamada *clase exponencial*. En virtud del Teorema anterior todo modelo de la clase exponencial admite un estadístico suficiente. Es fácil comprobar además, que si un modelo pertenece a la clase exponencial, entonces las distribuciones de la forma

$$p(\theta) \propto G(\theta)^\alpha \exp\{\beta\zeta(\theta)\} \quad (2)$$

constituyen una familia conjugada para aquellos pares de valores (α, β) que garanticen la convergencia de la integral $\int_{\Theta} p(\theta)d\theta$.

En efecto, si $p(z|\theta)$ es de la forma (1) y $p(\theta)$ de la forma (2) con parámetros α_0 y β_0 , la distribución final resulta ser

$$p(\theta|z) \propto p(z|\theta) p(\theta) \propto G(\theta)^{n+\alpha_0} \exp\{[t(z) + \beta_0]\zeta(\theta)\}$$

(*) Brown (1964) ofrece un análisis riguroso de las condiciones de regularidad que deben ser exigidas.

(**) Kendall & Stuart (Vol. 2, 1956/1977, p. 26) describen de forma elemental la estructura de la demostración.

que es también de la forma (2) con parámetros $\alpha_n = \alpha_0 + n$ y $\beta_n = \beta_0 + n(1 - \theta)$. En el caso de Bernoulli $G(\theta) = (1 - \theta)^\alpha \theta^\beta$ y $\xi(\theta) = \log\{\theta/(1 - \theta)\}$ de forma que sustituyendo en (2) la familia conjugada resulta ser del tipo

$$p(\theta) \propto (1 - \theta)^\alpha \left(\frac{\theta}{1 - \theta} \right)^\beta = (1 - \theta)^{\alpha - \beta} \theta^\beta$$

Claramente, para $\beta + 1 > 0$ y $\alpha > \beta - 1$, que son los valores de α y β que garantizan la convergencia de $\int_0^1 p(\theta) d\theta$, se genera la familia de distribuciones Beta.

La gran mayoría de los modelos probabilísticos utilizados en la práctica pertenecen a la *clase exponencial* definida por (1) de forma que tienen tanto *estadístico suficiente* como *familia conjugada* de distribuciones. En el Teorema 6.2.2 se resumen, para algunos de ellos, su estadístico suficiente, las correspondientes distribuciones conjugadas y la relación que existe entre los parámetros iniciales y finales.

La demostración de algunos de estos resultados ha sido hecha en el texto; el resto pueden ser consultados, por ejemplo, en DeGroot (1970, cap. 9).

TEOREMA 6.2.2. Familias conjugadas

Modelo	Est. Suficiente	Familia Conjugada
Bernoulli $Br(x \theta)$	(r, n) $r = \sum x_i$	Beta, $Be(\theta \alpha_i, \beta_i)$ $\alpha_n = \alpha_0 + r$ $\beta_n = \beta_0 + n - r$
Poisson $Po(x \lambda)$	(r, n) $r = \sum x_i$	Gamma, $Gal(\lambda \alpha_i, \beta_i)$ $\alpha_n = \alpha_0 + r$ $\beta_n = \beta_0 + n$
Normal $N(x \mu, \sigma)$ σ conocido $b = 1/\sigma^2$	(\bar{x}, n) $\bar{x} = \sum x_i/n$	Normal, $N(\mu \mu_i, \sigma_i)$ $\sigma_i = 1/\sqrt{b_i}$ $b_n = b_0 + nb$ $\mu_n = (\mu_0 b_0 + n\bar{x})/b_n$
Normal $N(x \mu, \sigma)$ μ conocido $b = 1/\sigma^2$	(s^2, n) $r = \sum (x_i - \mu)^2/n$	Gamma, $Gal(b \alpha_i, \beta_i)$ $\alpha_n = \alpha_0 + n/2$ $\beta_n = \beta_0 + ns^2/2$
Exponencial $Ex(x \theta)$	(r, n) $r = \sum x_i$	Gamma, $Gal(\theta \alpha_i, \beta_i)$ $\alpha_n = \alpha_0 + n$ $\beta_n = \beta_0 + r$

Los resultados descritos en esta sección permiten un análisis rápido de datos correspondientes a un modelo de la familia exponencial *cuando es po-*

sible aproximar la información inicial por un elemento de la familia conjugada correspondiente. Hay muchas situaciones, sin embargo, en que la información inicial *no puede* ser adecuadamente aproximada por un miembro de esa familia (por ejemplo, cuando existen dos zonas disjuntas de valores muy probables). En tal caso, no hay más remedio que determinar una distribución inicial que describa tal información, aplicar el Teorema de Bayes, y realizar los cálculos necesarios para obtener las probabilidades que se necesitan recurriendo, si es necesario, a métodos de integración numérica mediante un ordenador.

Cuando la información inicial *si* que puede ser descrita por un miembro de la familia conjugada, resulta sencillo determinar sus parámetros mediante técnicas semejantes a las descritas en la Sección 5.1. En caso contrario, suele ser necesario recurrir al uso de un programa interactivo cuyas características dependerán, en general, del ordenador utilizado.

Ejemplo 6.2.1. Tiempos de espera

Se sabe que el tiempo en minutos que transcurre entre dos emisiones consecutivas de una materia radioactiva utilizada como trazador es una cantidad aleatoria con una distribución exponencial de parámetro θ . De acuerdo con la información de que se dispone, el valor más probable de tal parámetro θ es 0,5 y se está muy seguro (probabilidad 0,95) de que θ es mayor de 0,25. Se observan cuatro emisiones consecutivas y los tiempos de espera entre ellos resultan ser 2,4, 2 y 1,9 minutos. Determinar la probabilidad de que el valor de θ sea mayor de 0,5.

Por hipótesis,

$$p(x|\theta) = Ex(x|\theta) = Gal(x|1, \theta) = \theta e^{-\theta x}, \quad x > 0, \theta > 0$$

La información inicial sobre θ es unimodal (con moda 0,5); podemos pues intentar describirla mediante un miembro de la correspondiente familia conjugada que, en este caso (Teorema 6.2.2) es la familia de distribuciones Gamma. Si $p(\theta) = Gal(\theta|\alpha, \beta)$, el valor más probable de θ es (Teorema 4.5.3) $(\alpha - 1)/\beta$. Tenemos, pues, que resolver el sistema

$$\begin{cases} (\alpha - 1)/\beta = 0,5 \\ \int_{0,25}^{\infty} Gal(\theta|\alpha, \beta) = 0,95 \end{cases}$$

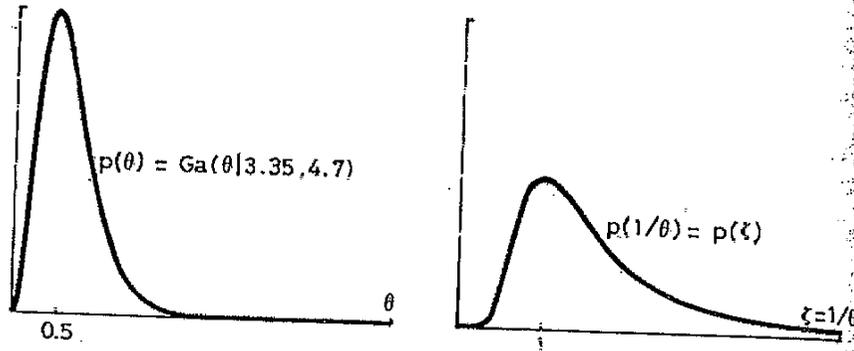
que puede escribirse en la forma

$$\begin{cases} \beta = 2\alpha - 2 \\ 1 - \int_0^{0,25} Gal(\theta|\alpha, \beta) = 0,95 \end{cases}$$

es decir

$$F_{\alpha, 2\alpha-2}(0,25) = 0,05$$

dónde $F_{\alpha, 2\alpha-2}$ es la función de distribución de una cantidad aleatoria Gamma con parámetros α y $2\alpha - 2$. Utilizando unas Tablas (o recurriendo a una aproximación normal para log θ : ver Ejemplo 6.3.1) se obtiene $\alpha = 3,35$, de forma que la distribución es $p(\theta) = Ga(\theta|3,35, 4,7)$. Puesto que $E(x) = 1/\theta$, la representación de la correspondiente distribución $1/\theta$ puede ayudar a entender el significado de esta distribución inicial, puesto que describe la información de que inicialmente se dispone sobre el tiempo *medio* que transcurre entre dos emisiones consecutivas.



De acuerdo con el Teorema 4.4.1, si $\zeta(\theta) = 1/\theta$

$$p(\zeta) = p(\theta) / |\partial\zeta/\partial\theta| \propto \theta^{2.35} \exp(-4,7\theta) = \zeta^{-4,35} \exp(-4,7/\zeta)$$

cuya representación gráfica es la de la figura.

Si la gráfica de la distribución inicial encontrada se juzga consistente con la información inicial, podemos proceder a determinar la correspondiente distribución final. Para el modelo exponencial, el estadístico suficiente es $(r = \sum x_i, n)$; en nuestro caso, $r = 6,3$ y $n = 3$. La distribución final correspondiente es pues, haciendo uso del Teorema 6.2.2, $p(\theta|z) = G(\theta|\alpha_n, \beta_n)$ con $\alpha_n = \alpha_0 + n = 3,35 + 3 = 6,35$ y $\beta_n = \beta_0 + r = 4,7 + 6,3 = 11$. En consecuencia, la probabilidad pedida será

$$p(\theta > 0,5|z) = \int_{0,5}^{\infty} Ga(\theta|6,35, 11) d\theta \approx 0,565$$

El valor de la integral puede obtenerse mediante Tablas o bien, como veremos en el Ejemplo 6.3.2, recurriendo a una aproximación normal para log θ .

6.3. Aproximación normal a la distribución final

La mayor parte de las densidades de probabilidad que aparecen como resultado de problemas reales de inferencia son difíciles de integrar para obtener las probabilidades deseadas. Aunque, naturalmente, puede recurrirse a la integración numérica o a la preparación de tablas, resulta conveniente disponer de métodos que permiten obtener resultados aproximados con un

mínimo de equipo. Puesto que, a diferencia de lo que sucede con las distribuciones Beta o Gamma (que requieren una tabla para cada par de valores de sus parámetros), todas las probabilidades que se derivan de distribuciones normales pueden obtenerse a partir de una única tabla (la de la normal *standard*) resulta muy conveniente disponer de métodos que permitan reducir el cálculo de probabilidades asociadas a una distribución cualquiera al cálculo de probabilidades asociadas a una distribución normal.

Para poder hablar de *aproximación de distribuciones* con un mínimo de rigor necesitamos definir una *medida* de la *discrepancia* existente entre una densidad de probabilidad y su aproximación.

DEFINICIÓN 6.3.1. La discrepancia entre una densidad de probabilidad $p(\theta)$ y su aproximación $q(\theta)$ es el valor de la integral

$$\delta\{p(\theta), q(\theta)\} = \int p(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta$$

Si θ tiene una distribución discreta basta, como siempre, sustituir en la definición las densidades de probabilidad por probabilidades y la integral por una suma.

La aproximación $q(\theta)$ será tanto mejor cuanto *menor* sea la discrepancia $\delta\{p(\theta), q(\theta)\}$. La Definición 6.3.1 puede justificarse con un argumento de tipo axiomático (Bernardo, 1980 b); por otra parte, recordando la Definición 5.4.1, su significado intuitivo es claro: $\delta\{p(\theta), q(\theta)\}$ es la cantidad de información sobre θ que resulta necesaria para pasar de la aproximación $q(\theta)$ a la densidad de probabilidad verdadera $p(\theta)$.

Ejemplo 6.3.1. Aproximación Poisson de una Binomial

Representar en función de θ , y para los valores $n = 1, 2$ y 10 , la discrepancia $\delta\{Bi(x|n, \theta), Po(x, n\theta)\}$ entre una distribución Binomial $Bi(x|n, \theta)$ y su aproximación $Po(x|n\theta)$ (Teorema 4.2.1).

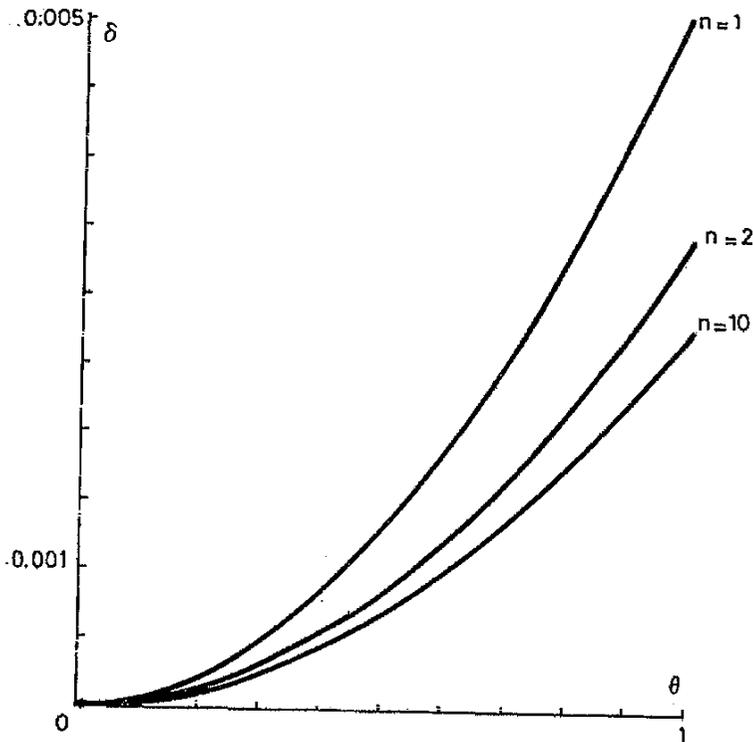
Se trata de representar la función

$$\delta = f(\theta|n) = \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$$

$$p_i = \binom{n}{i} \theta^i (1-\theta)^{n-i}$$

$$q_i = e^{-n\theta} (n\theta)^i / i!$$

para los valores pedidos de n . El resultado es



Como podía esperarse, el valor de δ decrece cuando n aumenta y/o cuando θ disminuye. La observación de la gráfica sugiere que para obtener buenas aproximaciones Poisson de distribuciones bihomiales, la condición de que θ tenga un valor pequeño tiene más importancia que la condición de que n tenga un valor grande.

TEOREMA 6.3.1. *La mejor aproximación normal a una distribución $p(\theta)$, es $N\{\theta|E(\theta), D(\theta)\}$, esto es aquella distribución normal con la misma media y desviación típica que la distribución original.*

Demostración

Nos limitaremos a señalar sus líneas esenciales. La discrepancia entre $p(\theta)$ y una distribución normal cualquiera $N(\theta|\mu, \sigma)$ viene dada por

$$\delta(p(\theta), N(\theta|\mu, \sigma)) = \int p(\theta) \log \frac{p(\theta)}{N(\theta|\mu, \sigma)} d\theta$$

Puede comprobarse, por derivación bajo el signo integral, que el valor mínimo de δ se alcanza cuando $\mu = \int \theta p(\theta) d\theta = E(\theta)$ y $\sigma^2 = \int (\theta - E(\theta))^2 p(\theta) d\theta = D^2(\theta)$.

Es intuitivamente inmediato que, en general, resulta imposible aproximar bien, esto es con una discrepancia pequeña, una distribución cualquiera mediante una distribución normal; lo que resulta sin embargo frecuentemente posible es encontrar una transformación monótona de θ , $\zeta = \zeta(\theta)$, razonablemente sencilla para que el problema sea tratable, cuya densidad de probabilidad $p(\theta)$ sea aproximadamente normal; esto resulta suficiente para resolver nuestro problema.

En efecto, si deseamos calcular $p[a < \theta < b]$ y sabemos que la función monótona de θ , $\zeta = \zeta(\theta)$, tiene una distribución aproximadamente normal $p(\zeta) = N\{\zeta|E(\zeta), D(\zeta)\}$, tenemos

$$p[a < \theta < b] = p[\zeta(a) < \zeta < \zeta(b)]$$

de forma que, utilizando la Ecuación 4.3.1, la probabilidad buscada resulta ser

$$\Phi \left\{ \frac{\zeta(b) - E(\zeta)}{D(\zeta)} \right\} - \Phi \left\{ \frac{\zeta(a) - E(\zeta)}{D(\zeta)} \right\}$$

donde Φ es la función de distribución de la normal standard.

Para una determinada distribución inicial $p(\theta)$, el problema de encontrar la mejor transformación normalizadora posible $\zeta = \zeta(\theta)$ en el sentido de minimizar la discrepancia entre $p(\theta)$ y su mejor aproximación normal equivale, en virtud de los Teoremas 4.4.1 y 6.3.1 a encontrar la función monótona de θ , $\zeta = \zeta(\theta)$ que minimiza la integral

$$\int p(\zeta) \log [p(\zeta)/N\{\zeta|E(\zeta), D(\zeta)\}] d\zeta \tag{1}$$

donde $p(\zeta) = p(\theta)/|\partial\zeta/\partial\theta|$ y $E(\zeta), D(\zeta)$ son respectivamente la media y la desviación típica de ζ . Se trata de un problema difícil de cálculo de variaciones para el que sólo se conocen resultados parciales; algunos de estos resultados están contenidos en los Teoremas 6.3.2 y 6.3.3, cuya demostración (Bernardo, 1980 a) omitiremos.

TEOREMA 6.3.2. *Si $p(\theta) = Be\{\theta|\alpha, \beta\}$, entonces $\zeta(\theta) = \log\{\theta/(1-\theta)\}$ es la mejor transformación normalizadora dentro de una amplia clase de transformaciones, y se verifica que $p(\zeta) = N\{\zeta|E(\zeta), D(\zeta)\}$ con*

$$E(\zeta) = \Psi(\alpha) - \Psi(\beta) = \log(\alpha/\beta) + (\alpha - \beta)/(2\alpha\beta)$$

$$D^2(\zeta) = \Psi'(\alpha) + \Psi'(\beta) = (\alpha + \beta)/\alpha\beta$$

donde $\Psi(x)$ es la función digamma (*) y las aproximaciones son válidas para valores no muy pequeños de α y β .

Este resultado había sido parcialmente anticipado en los Ejemplos 4.4.2 y 4.5.3, y utilizado en los Ejemplos 5.1.4 y 5.4.2 para obtener aproximadamente determinadas probabilidades.

TEOREMA 6.3.3. Si $p(\theta) = Ga(\theta|\alpha, \beta)$, entonces $\zeta(\theta) = \log \theta$ es la mejor transformación normalizadora dentro de una amplia clase de transformaciones, y se verifica que $p(\zeta) \approx N\{\zeta|E(\zeta), D(\zeta)\}$ con

$$E(\zeta) = \Psi(\alpha) - \log \beta \approx \log(\alpha/\beta) - 1/(2\alpha)$$

$$D^2(\zeta) = \Psi'(\alpha) \approx 1/\alpha$$

donde $\Psi(x)$ es la función digamma y las aproximaciones son válidas para valores no muy pequeños de α y β .

Ejemplo 6.3.2. Ajuste de distribuciones gamma

Determinar aproximadamente aquella distribución Gamma de una cantidad aleatoria θ cuya moda sea 0,5 y tal que $P[\theta < 0,25] = 0,05$.

Puesto que $(\alpha - 1)/\beta$ es la moda de una distribución Gamma, tenemos $(\alpha - 1)/\beta = 0,5$ y, por tanto, $\beta = 2\alpha - 2$. Por otra parte, si $\zeta = \log \theta$ tenemos, en virtud del Teorema 6.3.3,

$$\begin{aligned} P[\theta < 0,25] &= P[\log \theta < \log 0,25] = \\ &= P[\zeta < \log 0,25] = \Phi\{[\log 0,25 - E(\zeta)]/D(\zeta)\} = 0,05 \end{aligned}$$

con $E(\zeta) = \log\{\alpha/(2\alpha - 2)\} - 1/2\alpha$ y $D(\zeta) \approx 1/\sqrt{2\alpha}$. En consecuencia, utilizando las Tablas de la distribución Normal,

$$[\log 0,25 - \log\{\alpha/(2\alpha - 2)\} + 1/2\alpha]/\sqrt{2\alpha} = -1,6449$$

ecuación cuya solución aproximada, calculada por prueba y error, es $\alpha = 3,35$.

Ejemplo 6.3.3. Cálculo de probabilidades en distribuciones gamma

Si la información de que se dispone sobre una cantidad aleatoria θ puede describirse mediante la distribución $Ga(\theta|6,35, 11)$, determinar la probabilidad de que θ sea mayor que 0,5.

(*) La función digamma se define mediante la expresión $\Psi(x) = \Gamma'(x)/\Gamma(x)$ donde $\Gamma(x)$ es la función gamma, ya definida. Puede demostrarse que, para todo x , $\Psi(x+1) = \Psi(x) + 1/x$ con $\Psi(1) = -\gamma$ y $\Psi(1) = -\gamma - 2 \log 2$ donde $\gamma \approx 0,5772$ es la constante de Euler. Análogamente, $\Psi'(x+1) = \Psi'(x) - 1/x^2$, con $\Psi'(1) = \pi^2/6$ y $\Psi'(1/2) = \pi^2/2$.

Utilizando de nuevo el Teorema 6.3.3 con $\zeta(\theta) = \log(\theta)$, tenemos

$$p[\theta < 0,5] = P[\log \theta < \log 0,5] = \Phi\{[\log 0,5 - E(\zeta)]/D(\zeta)\}$$

donde

$$E(\zeta) = \Psi(6,35) - \log 11 = -0,628$$

$$D^2(\zeta) = \Psi'(6,35) = 0,157$$

de forma que

$$p[\theta < 0,5] = \Phi(-0,164) = 1 - \Phi(0,164) = 0,435$$

6.4. Comportamiento asintótico de la distribución final

Como demostramos en el Capítulo 5 y se ha ilustrado repetidamente con numerosos ejemplos, la distribución final combina la información proporcionada por los datos con la información de que inicialmente se dispone.

Frecuentemente, la información proporcionada por los datos es comparativamente muy importante debido a la existencia de abundantes datos experimentales; en tales casos, la distribución final admite una aproximación asintótica que es tanto más precisa cuanto mayor es la cantidad de datos de que se dispone. Más concretamente, para casi todos los modelos probabilísticos $p(x|\theta)$ y cualquiera que sea la distribución inicial $p(\theta)$ de una cantidad aleatoria, su distribución final $p(\theta|z)$ después de observar los resultados experimentales $z = \{x_1, \dots, x_n\}$ se va aproximando a una distribución normal a medida que crece el tamaño n de la muestra. Este es el contenido fundamental del resultado siguiente.

TEOREMA 6.4.1. Sean x los resultados de un experimento e cuya distribución es $p(x|\theta)$ y supongamos que el conjunto X de valores posibles de x no depende de θ (*); sea $p(\theta)$ la distribución inicial de θ y sean $z = \{x_1, \dots, x_n\}$ los resultados de n realizaciones independientes de tal experimento. Entonces, para valores de n suficientemente grandes, la distribución final de θ , $p(\theta|x_1, \dots, x_n)$ es aproximadamente normal con media $E(\theta|z)$ y desviación típica $D(\theta|z)$ dadas por

$$E(\theta|z) = \{\theta_0 b_0 + \theta h(\theta)\} / \{b_0 + h(\theta)\}$$

$$D(\theta|z) = 1 / \{\sqrt{b_0} + h(\theta)\}$$

donde θ es el máximo de la función de verosimilitud $p(z|\theta) = \prod p(x_i|\theta)$, θ_0 es la moda de la distribución inicial de θ ,

(*) Estrictamente, es necesario exigir algunas condiciones matemáticas adicionales para que el Teorema sea cierto. Resulta, sin embargo, que en casi todos los problemas que aparecen en la práctica se cumplen tales condiciones (Johnson, 1967, 1970).

$$b(\theta) = - \left. \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta) \right|_{\theta=\hat{\theta}}$$

y, finalmente,

$$b_0 = - \left. \frac{\partial^2}{\partial \theta^2} \log p(\theta) \right|_{\theta=\theta_0}$$

Demostación

No es posible dar una demostración rigurosa de este Teorema al nivel matemático elegido para este libro; sin embargo, si que resulta posible justificar intuitivamente el resultado. En efecto, en virtud del Teorema de Bayes

$$\begin{aligned} p(\theta|x_1, \dots, x_n) &\propto \prod_{i=1}^n p(x_i|\theta)p(\theta) \\ &= \exp \left\{ \sum_{i=1}^n \log p(x_i|\theta) + \log p(\theta) \right\} \end{aligned} \quad (1)$$

Si denotamos por $L_n(\theta) = \sum \log p(x_i|\theta)$ al logaritmo de la función de verosimilitud y lo desarrollamos en serie alrededor de su máximo $\hat{\theta}$, podemos escribir, para n grande,

$$L_n(\theta) = L_n(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^2 \left. \frac{\partial^2}{\partial \theta^2} L_n(\theta) \right|_{\theta=\hat{\theta}}$$

puesto que la primera derivada de $L_n(\theta)$ calculada en $\hat{\theta}$ será nula por tratarse de un máximo. Por otra parte, desarrollando $\log p(\theta)$ alrededor de su moda θ_0 obtenemos igualmente

$$\log p(\theta) = \log p(\theta_0) + \frac{1}{2} (\theta - \theta_0)^2 \left. \frac{\partial^2}{\partial \theta^2} \log p(\theta) \right|_{\theta=\theta_0}$$

de forma que sustituyendo en (1) y omitiendo innecesarias constantes de proporcionalidad

$$p(\theta|x_1, \dots, x_n) \propto \exp \left\{ - \frac{b(\theta)}{2} (\theta - \hat{\theta})^2 - \frac{b_0}{2} (\theta - \theta_0)^2 \right\} \quad (2)$$

donde

$$b(\theta) = - \left. \frac{\partial^2}{\partial \theta^2} L_n(\theta) \right|_{\theta=\hat{\theta}} = - \sum_{i=1}^n \left. \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta) \right|_{\theta=\hat{\theta}}$$

$$b_0 = - \left. \frac{\partial^2}{\partial \theta^2} \log p(\theta) \right|_{\theta=\theta_0}$$

Finalmente, utilizando la identidad 5.4.14, la ecuación (2) puede escribirse

$$p(\theta|x_1, \dots, x_n) \propto \exp \left\{ - \frac{b_0 + b(\theta)}{2} (\theta - \theta_n)^2 \right\}$$

donde $\theta_n = \{\theta_0 b_0 + \hat{\theta} b(\hat{\theta})\} / \{b_0 + b(\hat{\theta})\}$ lo que, comparando con la definición de la distribución final, constituye el resultado buscado.

Analizaremos a continuación el resultado de aplicar el Teorema anterior a dos de los modelos probabilísticos más frecuentes.

Consideremos primero el modelo de Bernoulli. Sean $x = \{x_1, \dots, x_n\}$ los resultados de n observaciones independientes de una distribución de Bernoulli con parámetro θ , esto es

$$p(x|\theta) = \prod p(x_i|\theta) = \theta^r (1 - \theta)^{n-r}, \quad r = \sum x_i$$

y supongamos que la información inicial de que se dispone sobre θ puede describirse mediante una distribución normal $N(\theta|\theta_0, \sigma_0)$. En este caso, el valor más probable de θ antes de realizar el experimento es obviamente θ_0 ; además, con la notación del Teorema 6.4.1,

$$b_0 = - \left. \frac{\partial^2}{\partial \theta^2} \log N(\theta|\theta_0, \sigma_0) \right|_{\theta=\theta_0} = \frac{1}{\sigma_0^2} \quad (3)$$

El valor de θ que maximiza $p(x_n|\theta)$ resulta ser $\hat{\theta} = r/n$ y, con $r = n\hat{\theta}$,

$$b(\hat{\theta}) = - \left. \frac{\partial^2}{\partial \theta^2} \sum_{i=1}^n \log p(x_i|\theta) \right|_{\theta=\hat{\theta}} = \frac{n}{\hat{\theta}(1-\hat{\theta})} = \frac{n^2}{r(n-r)} \quad (4)$$

En consecuencia, en virtud del Teorema 6.4.1, la distribución final de θ es, para valores grandes de n , aproximadamente $N\{\theta|E(\theta|x), D(\theta|x)\}$, con

$$E(\theta|x) = \{\theta_0 b_0 + \hat{\theta} b(\hat{\theta})\} / \{b_0 + b(\hat{\theta})\} \quad (5)$$

$$D(\theta|x) = 1 / \sqrt{\{b_0 + b(\hat{\theta})\}} \quad (6)$$

donde b_0 y $b(\hat{\theta})$ vienen dados por (3) y (4).

Si la distribución inicial hubiese sido $Be(\theta|\alpha, \beta)$, entonces el valor de θ inicialmente más probable hubiese sido la moda de esa distribución, $\theta_0 = (\alpha - 1) / (\alpha + \beta - 2)$ y el valor de b_0 hubiese resultado ser

$$b_0 = - \left. \frac{\partial^2}{\partial \theta^2} \log Be(\theta|\alpha, \beta) \right|_{\theta=\theta_0} = \frac{(\alpha + \beta - 2)^2}{(\alpha - 1)(\beta - 1)} \quad (7)$$

Ejemplo 6.4.1. Probabilidad de contagio

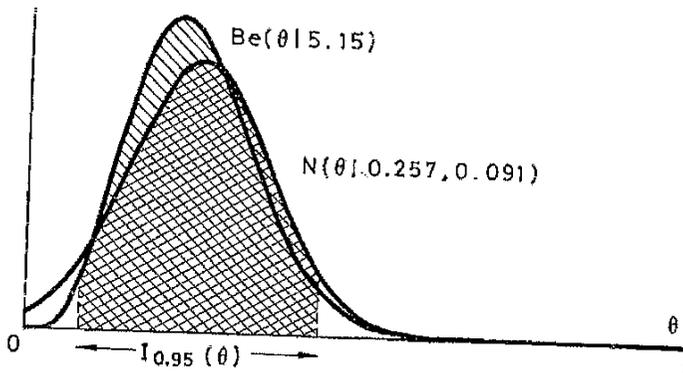
La probabilidad de que una persona sometida a unas condiciones bien definidas contraiga una enfermedad si está entre 0,078 y 0,436 con probabilidad 0,95. Se observa que de 60 personas sometidas a tales condiciones 12 han contraído la enfermedad. Determinar aproximadamente la región creíble final de nivel 0,95 correspondiente a la probabilidad investigada, según se suponga que la distribución inicial es de tipo normal o de tipo Beta.

a) Sea $p(\theta) = N(\theta|\theta_0, \sigma_0)$. Por hipótesis, $p(0,078 < \theta < 0,436) = 0,95$ y por tanto, dada la simetría de la normal, $p(\theta < 0,436) = 0,975$ y $\theta_0 = (0,078 + 0,436)/2 = 0,257$. Consecuentemente

$$p[\theta < 0,436] = p\left[\frac{\theta - \theta_0}{\sigma_0} < \frac{0,436 - \theta_0}{\sigma_0}\right] = \Phi\left(\frac{0,179}{\sigma_0}\right) = 0,975$$

y por tanto, $0,179/\sigma_0 = 1,96$, $\sigma_0 = 0,091$ y $b_0 = 1/\sigma_0^2 = 119,9$. Por otra parte, $\theta = r/n = 12/60 = 0,2$ y $h(\theta) = n/\{\theta(1-\theta)\} = 375$ de forma que, en virtud de (3) y (4), $E(\theta|z) = 0,214$ y $D(\theta|z) = 0,045$. Finalmente, utilizando la Ecuación 6.1.3 el intervalo de confianza pedido resulta ser $I_{0,95}(\theta|z) = (0,126, 0,302)$.

b) Sea $p(\theta) = Be(\theta|\alpha, \beta)$. Utilizando Tablas, o haciendo uso de la aproximación normal estudiada en la Sección anterior, resulta que $Be(\theta|5, 15)$ es la distribución Beta cuyo intervalo inicial de nivel 0,95 es (0,078, 0,436). En consecuencia, $\theta_0 = (\alpha - 1)/(\alpha + \beta - 2) = 0,222$ y, utilizando (7), $b_0 = 104,1$. Haciendo uso de (3) y (4) para combinar estos valores con los de θ y $h(\theta)$ anteriormente obtenidos, resulta $E(\theta|z) = 0,205$ y $D(\theta|z) = 0,046$; el intervalo final de confianza es ahora $I_{0,95}(\theta|z) = (0,115, 0,294)$, no muy lejos del anterior.



La diferencia entre las hipótesis a) y b) sobre la forma de la distribución inicial puede apreciarse en la figura.

Consideremos ahora el modelo normal. Supongamos que $z = \{x_1, \dots, x_n\}$

son los resultados de n observaciones independientes $N(x_i|\mu, \sigma)$, con σ conocido, y supongamos que la información inicial sobre μ puede describirse mediante una distribución normal $N(\mu|\mu_0, \sigma_0)$. Como en el caso anterior, el valor inicialmente más probable de μ será μ_0 y, con la notación del Teorema 6.4.1, $b_0 = 1/\sigma_0^2$.

Por otra parte, es fácil comprobar que el valor que maximiza $p(z|\mu)$ es $\hat{\mu} = \bar{x}$ y que $h(\hat{\mu}) = n/\sigma^2$. En consecuencia, en virtud del Teorema 6.4.1, la distribución asintótica de μ es Normal con media y desviación típica dadas por

$$E(\mu|z) = (b_0\mu_0 + n\bar{x}h)/(b_0 + nb)$$

$$D(\mu|z) = 1/(b_0 + nb)$$

donde $b_0 = 1/\sigma_0^2$ y $b = 1/\sigma^2$. Pero éstos son precisamente los parámetros μ_n y σ_n obtenidos (Ecuaciones 5.4.18 a 5.4.20) al calcular exactamente la distribución final de μ en las condiciones descritas. En consecuencia, si el modelo es normal $N(x|\mu, \sigma)$ con desviación típica conocida y la distribución inicial es normal $N(\mu|\mu_0, \sigma_0)$, el Teorema 6.4.1 no se limita a dar una aproximación a la distribución final sino que da lugar al resultado exacto.

El Teorema 6.4.1 tiene, obviamente una aplicación inmediata para resolver de forma aproximada problemas de inferencia en los que el número de observaciones experimentales es suficientemente grande. En la próxima sección nos resultará necesario un importante corolario del Teorema 6.4.1:

TEOREMA 6.4.2. Si, en las condiciones del Teorema 6.4.1, el tamaño n de la muestra es muy grande, la distribución final $p(\theta|z)$ puede ser aproximada por

$$p^*(\theta|\theta) = N\{\theta|\theta, D(\theta)\}$$

donde θ es el estimador máximo verosímil de θ , $D(\theta) = \{b(\theta)\}^{-1/2}$ y $b(\theta)$ viene definido por

$$b(\theta) = - \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta) \Big|_{\theta=\theta}$$

Además, $b(\theta)$ crece linealmente con n , de forma que cuando n tiende a infinito, $D(\theta)$ tiende a cero y θ tiende a θ .

En efecto, cuando el tamaño n de la muestra es muy grande, y con la notación del Teorema 6.4.1, b_0 es muy pequeño frente a $b(\theta)$ y puede susti-

tirse por cero sin grandes alteraciones en el resultado final. La segunda parte del Teorema es una consecuencia del *Teorema central del límite* (Sección 4.3) aplicado a las cantidades aleatorias $\partial^2 \log p(x_i|\theta)/\partial\theta^2$. Una consecuencia importante de esta segunda parte es que, para valores grandes de n , un parámetro desconocido θ puede ser aproximado por su estimador máximo-verosímil $\hat{\theta}$. Técnicamente, se dice que una función $\theta = \theta(x)$ de los datos es un *estimador consistente* de θ si converge a θ cuando $n \rightarrow \infty$; el estimador máximo-verosímil $\hat{\theta}$ es un ejemplo de estimador consistente.

La distribución final $p^*(\theta|\hat{\theta})$ obtenida en el Teorema 6.4.2 es la que generalmente se conoce como *distribución asintótica* de θ correspondiente al modelo $p(x|\theta)$. Se ha elegido este nombre debido al hecho de que *cualquiera que sea la distribución inicial* $p(\theta)$, la verdadera distribución final $p(\theta|x)$ se aproxima a $p^*(\theta|\hat{\theta})$ cuando n tiende a infinito de forma parecida a como una curva se aproxima a su asíntota.

Claramente, la distribución asintótica no depende más que del modelo probabilístico y de los datos. Más aún, puede observarse que sólo depende de los datos a través del estimador máximo-verosímil $\hat{\theta}$. Técnicamente, se dice que una función $t = t(x)$ de los datos es un *estadístico asintóticamente suficiente* para θ si la distribución asintótica de θ solo depende de los datos a través de t , de forma que, para muestras grandes, un estadístico asintóticamente suficiente resume casi toda la información relevante contenida en los datos; el estimador máximo-verosímil $\hat{\theta}$ es un ejemplo de estadístico asintóticamente suficiente.

El Teorema 6.4.2 puede generalizarse al caso en que el modelo, $p(x|\theta_1, \dots, \theta_k)$ dependa de varios parámetros de forma que $\theta = \{\theta_1, \dots, \theta_k\}$ es un vector aleatorio.

TEOREMA 6.4.3. *La distribución asintótica de un vector aleatorio $\theta = \{\theta_1, \dots, \theta_k\}$ de dimensión k es la distribución normal k -variante $p^*(\theta|\hat{\theta}) = N\{\theta|\hat{\theta}, H(\hat{\theta})^{-1}\}$ donde $\hat{\theta}$ es el vector que maximiza la función de verosimilitud $p(x|\theta)$ y la matriz de precisión $H(\hat{\theta})$ tiene como elemento genérico*

$$h_{ij}(\hat{\theta}) = - \sum_{i=1}^k \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x_i|\hat{\theta}) \Big|_{\theta=\hat{\theta}}$$

Comentaremos, a continuación, la distribución asintótica conjunta de los parámetros de una distribución normal así como sus correspondientes distribuciones marginales y condicionales.

De acuerdo con el Teorema 6.4.3, se trata de una distribución normal bivalente con media

$$\theta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix} = \begin{pmatrix} \bar{x} \\ s \end{pmatrix}$$

donde $\bar{x} = \Sigma x_i/n$ y $s^2 = \Sigma(x_i - \bar{x})^2/n$ y con matriz de precisión

$$H(\hat{\theta}) = H(\mu, \sigma) = \begin{pmatrix} n/s^2 & 0 \\ 0 & 4n/s^2 \end{pmatrix}$$

Puesto que se trata de una matriz diagonal, los parámetros μ y σ resultan ser *asintóticamente independientes* y, como consecuencia, sus distribuciones asintóticas marginales y condicionales coinciden y son (Teorema 4.6.7) de la forma

$$p^*(\mu|\bar{x}, s, \sigma) = p^*(\mu|\bar{x}, s) \approx N\{\mu|\bar{x}, s/\sqrt{n}\} \quad (8)$$

$$p^*(\sigma|\bar{x}, s, \mu) = p^*(\sigma|\bar{x}, s) \approx N\{\sigma|s, s/(2\sqrt{n})\} \quad (9)$$

Estos resultados, y particularmente el primero, son de gran importancia práctica.

Ejemplo 6.4.2. Composición del líquido cefalorraquídeo

Determinar *aproximadamente* la probabilidad de que la cantidad de potasio en el líquido cefalorraquídeo de una persona supere los 11 mg/100 ml sabiendo que la media de las cantidades observadas en 900 personas escogidas al azar es 11,57 y su desviación típica 11,76 mg/100 ml.

Puesto que disponemos de mucha información experimental, podemos, en primera aproximación, prescindir de cualquier información inicial y determinar, utilizando el resultado anterior, la distribución asintótica del valor medio μ de la cantidad de potasio en el líquido cefalorraquídeo.

De acuerdo con la ecuación (8), la distribución asintótica de μ cuando σ es desconocido es $N(\mu|\bar{x}, s/\sqrt{n})$. Utilizando la ecuación 5.6.4 y aproximando σ por su estimador máximo-verosímil s puesto que el tamaño muestral lo permite,

$$p^*(x|\bar{x}, s) = N\{x|\bar{x}, \sqrt{(s^2 + s^2/n)}\} = N\{x|\bar{x}, s\sqrt{1 + 1/n}\}$$

En nuestro ejemplo, $\bar{x} = 11,57$ y $s\sqrt{1 + 1/n} = 1,761$ y, por tanto, la probabilidad pedida es, aproximadamente,

$$p\{x > 11|\bar{x}, s\} = 1 - \Phi\left(\frac{11 - 11,57}{1,761}\right) = \Phi(0,324) = 0,627$$

6.5. Distribuciones finales de referencia

La inferencia estadística Bayesiana ha sido repetidamente descrita como la metodología que permite combinar de forma consistente la información inicial con la información experimental. Tal descripción sugiere inmediatamente un problema importante que todavía carece de una respuesta incontrovertida: se trata de determinar la forma en que deben realizarse inferencias cuando no se dispone de información inicial o cuando tal información no se quiere o no se puede utilizar.

Desde un punto de vista Bayesiano, el problema quedaría resuelto si pudiésemos determinar una *distribución inicial de referencia*, que describa la situación en que los datos experimentales contienen *toda* la información relevante, en lugar de proporcionar tan sólo una parte de ella como sucede cuando se dispone de información inicial. Para determinar tal distribución inicial utilizaremos el concepto de información esperada de un experimento (Definición 5.4.2).

Supongamos, pues, que pretendemos realizar inferencias sobre el valor de θ mediante la realización de un experimento ε cuyos resultados z tienen una distribución $p(z|\theta)$. En tal caso, si $p(\theta)$ es la distribución inicial de θ , $I^0\{\varepsilon, p(\theta)\}$, definido en la Sección 5.4, mide la cantidad de información que puede esperarse del experimento ε sobre el valor de θ . Sea $\varepsilon(k)$ el experimento que consiste en k realizaciones sucesivas e independientes del experimento ε , y considérese la cantidad $I^0\{\varepsilon(k), p(\theta)\}$, es decir la cantidad de información sobre el valor de θ que puede esperarse de k repeticiones de ε , cuando la distribución inicial de θ es $p(\theta)$. Repitiendo indefinidamente el experimento, se llegaría a conocer *exactamente* el valor de θ ; en consecuencia,

$$\lim_{k \rightarrow \infty} I^0\{\varepsilon(k), p(\theta)\}$$

mide, para cada $p(\theta)$, la *cantidad de información desconocida* sobre θ cuando la distribución inicial de θ es $p(\theta)$. Resulta entonces natural definir la distribución inicial de referencia como aquella distribución $\pi(\theta)$ que hace máxima la información inicialmente desconocida. La distribución final de referencia se obtiene entonces mediante el simple uso del Teorema de Bayes. En las páginas siguientes procederemos a desarrollar estas ideas con cierto detalle. El lector interesado únicamente en los resultados finales puede omitir el resto de esta sección y consultar directamente el Teorema 6.5.4.

DEFINICIÓN 6.5.1. Sea \mathcal{C} la clase de distribuciones iniciales admisibles; esto es compatible con las características del problema. Sea ε un experimento cuyos resultados z tienen una distribución $p(z|\theta)$ y sea $I^0\{\varepsilon(k), p(\theta)\}$ la información que podría esperarse de k repeticiones independientes de ε cuando

la distribución inicial es $p(\theta)$. Sea $\pi_k(\theta)$ la distribución de θ que maximiza en \mathcal{C} el valor de $I^0\{\varepsilon(k), p(\theta)\}$; la distribución inicial de referencia es entonces

$$\pi(\theta) = \lim_{k \rightarrow \infty} \pi_k(\theta)$$

la distribución final de referencia, una vez observado el resultado z del experimento ε , es entonces $\pi(\theta|z) \propto p(z|\theta) \pi(\theta)$.

La obtención de las distribuciones iniciales de referencia y por tanto de sus correspondientes distribuciones finales es relativamente sencilla haciendo uso de los Teoremas 6.5.1 al 6.5.3 cuya demostración, basada en el comportamiento asintótico de la distribución final (Bernardo, 1979 b), omitiremos.

TEOREMA 6.5.1. Si θ solo puede tomar un número finito de valores $\{\theta_1, \dots, \theta_m\}$, entonces su distribución de referencia, cualquiera que sea el experimento ε , es la que maximiza en la clase \mathcal{C} de distribuciones iniciales admisibles el valor de la entropía.

$$H\{p(\theta)\} = - \sum_{i=1}^m p(\theta_i) \log p(\theta_i)$$

Si, en particular, \mathcal{C} es la clase de todas las distribuciones iniciales, entonces la distribución inicial de referencia es la distribución uniforme

$$\pi(\theta) = \{1/m, \dots, 1/m\}$$

Ejemplo 6.5.1. Diagnosis

Se sabe que un determinado paciente tiene una de las tres enfermedades θ_1, θ_2 o θ_3 , que la probabilidad p_2 de que tenga θ_2 no es mayor que 0,2 y que las probabilidades p_1 y p_3 de que tenga alguna de las otras dos enfermedades están comprendidas entre 0,1 y 0,6. Determinar la correspondiente distribución inicial de referencia, esto es, aquella que describe la información mencionada y *únicamente* esta información.

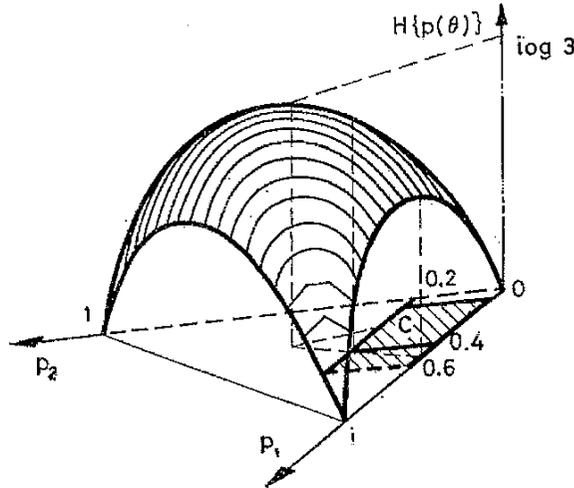
La clase \mathcal{C} de distribuciones admisibles está constituida por las distribuciones $\{(p_1, p_2, p_3), p_i \geq 0, \sum p_i = 1\}$ que satisfacen el sistema de ecuaciones

$$0,1 < p_1 < 0,6, \quad p_2 \leq 0,2, \quad 0,2 < p_3 < 0,6$$

Basta pues con determinar, entre los valores (p_1, p_2, p_3) que satisfacen tales condiciones, aquel que maximiza

$$H\{p(\theta)\} = -p_1 \log p_1 - p_2 \log p_2 - (1 - p_1 - p_2) \log (1 - p_1 - p_2)$$

La ecuación anterior es la de una superficie como la de la figura.



y su valor máximo en la clase \mathcal{C} se obtiene para los valores $p_1 = 0,4$, $p_2 = 0,2$, de forma que la distribución inicial pedida resulta ser $\pi(\theta) = \{0,4, 0,2, 0,4\}$.

El Teorema 6.5.1 proporciona, en el caso discreto y finito, un método sencillo para determinar la distribución de probabilidades $\pi(\theta)$ que describe la información sobre θ contenida en la definición de la clase \mathcal{C} de distribuciones admisibles, y únicamente tal información. En el próximo resultado, abordamos el caso continuo.

TEOREMA 6.5.2. Consideremos un experimento ε , la distribución de cuyos resultados z depende de un parámetro continuo θ y supongamos que su correspondiente distribución asintótica es $p^*(\theta|\theta)$. La distribución inicial de referencia para θ , correspondiente a ε , viene entonces dada por

$$\pi(\theta) = \lim_{k \rightarrow \infty} \exp \left\{ \int \dots \int p(z_1, \dots, z_k|\theta) \log p^*(\theta|\theta) dz_1, \dots, dz_k \right\}$$

donde θ es el valor de θ que maximiza $p(z_1, \dots, z_k|\theta)$.

Cuando la distribución final asintótica es normal, lo que según hemos visto en la sección anterior sucede muy frecuentemente, y resulta además que

no existen parámetros marginales, el Teorema anterior toma una forma especialmente sencilla.

TEOREMA 6.5.3. Supongamos que la distribución asintótica del parámetro θ de un experimento ε es $N\{\theta|\theta, D(\theta)\}$ donde θ es el estimador máximo verosímil de θ . Entonces, si \mathcal{C} es la clase de todas las distribuciones de θ que pueden definirse sobre un conjunto Θ , y no existen parámetros marginales, la distribución inicial de referencia para θ correspondiente al experimento ε es

$$\begin{aligned} \pi(\theta) &\propto 1/D(\theta), & \theta \in \Theta \\ &= 0, & \text{en su caso contrario} \end{aligned}$$

Utilizando el resultado del Teorema 6.4.2, la distribución inicial de referencia puede ser calculada a partir de la expresión

$$\begin{aligned} \pi(\theta) &\propto b(\theta)^{1/2} & \theta \in \Theta \\ &= 0, & \theta \notin \Theta \end{aligned} \tag{1}$$

donde la función $b(\theta)$ queda definida por la ecuación

$$b(\theta) = - \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta) \Big|_{\theta=\theta}$$

Frecuentemente, el cálculo de la distribución inicial de referencia no es más que un paso intermedio que resulta necesario para poder calcular, mediante el uso del Teorema de Bayes, la correspondiente distribución final de referencia. Intuitivamente, la distribución final de referencia, $\pi(\theta|z) \propto p(z|\theta) \pi(\theta)$, obtenida tras la observación de los resultados z del experimento ε definido por el modelo probabilístico $p(z|\theta)$ es aquella que refleja única y exclusivamente la información sobre θ proporcionada por el modelo y los datos, sin combinaria con la información inicial de que pueda disponerse sobre el valor de θ , más allá de la que se incorpore en la definición de la clase \mathcal{C} de distribuciones iniciales admisibles.

Consideraremos ahora con algún detalle el modelo de Bernoulli. En virtud del Teorema 6.4.2, la distribución asintótica del parámetro θ de una distribución de Bernoulli es $N\{\theta|\theta, D(\theta)\}$ con $\theta = r/n$, $r = \sum x_i$, $D(\theta) = b(\theta)^{-1/2}$, y

$$b(\theta) = - \sum \frac{\partial^2}{\partial \theta^2} \log p(x_i|\theta) \Big|_{\theta=\theta} = \frac{n}{\theta(1-\theta)}$$

En consecuencia, utilizando la ecuación (1), la distribución inicial de referencia para θ vendrá dada por

$$\pi(\theta) \propto b(\theta)^{1/2} \propto \theta^{-1/2} (1 - \theta)^{-1/2}$$

y, por tanto, comparando con la definición de una distribución Beta, $\pi(\theta) = Be(\theta|1/2, 1/2)$.

Consecuentemente, si se han realizado n observaciones independientes $\{x_1, \dots, x_n\}$ de ese modelo, la distribución final de referencia, que recoge la información así obtenida sobre el valor de θ , y únicamente esta información, será, cualquiera que sea la muestra, la que resulte de utilizar el Teorema de Bayes con $\pi(\theta) = Be(\theta|1/2, 1/2)$ como distribución inicial. Por tanto, haciendo uso del Teorema 6.2.2 para $\alpha_0 = \beta_0 = 1/2$, la distribución final de referencia resulta ser $\pi(\theta|r) = Be(\theta|r + 1/2, n - r + 1/2)$.

Ejemplo 6.5.2. Incidencia de una enfermedad rara

Con objeto de investigar la incidencia de una rara enfermedad se escogen al azar 1.500 individuos de la población bajo estudio y resulta que ninguno de ellos está afectado. Basándose únicamente en este resultado determinar la probabilidad de que una nueva persona elegida al azar entre las que no fueron incluidas en la muestra esté afectada por la enfermedad.

Si n es el tamaño de la muestra, r el número de personas afectadas y θ la proporción, obviamente desconocida, de personas afectadas en la población, la probabilidad de que una nueva persona tenga la enfermedad vendrá dada por

$$p(x = 1|r) = \int p(x = 1|\theta) p(\theta|r) d\theta = \int \theta p(\theta|r) d\theta = E(\theta|r)$$

es decir el valor esperado de la distribución final.

Puesto que únicamente disponemos de la información experimental, la distribución final de θ será la distribución de referencia, $Be(\theta|r + 1/2, n - r + 1/2)$, cuyo valor esperando (Teorema 4.5.3) es $(r + 1/2)/(n + 1)$. En nuestro caso, $n = 1.500$ y $r = 0$ de forma que la probabilidad pedida es $0,5/1.501 \approx 0,00033$.

Consideraremos ahora el modelo normal, con desviación típica conocida. Es inmediato demostrar, utilizando el Teorema 6.4.2 que la distribución asintótica del parámetro μ de una distribución normal $N(x|\mu, \sigma)$, con desviación típica conocida, σ , es $N(\mu|\bar{x}, \sigma/\sqrt{n})$, con $\mu = \bar{x}$. En consecuencia, en virtud del Teorema 6.5.2, la distribución inicial de referencia para μ debe cumplir

$$\pi(\mu) \propto 1/D(\mu) = \frac{\sqrt{n}}{\sigma}$$

Si suponemos que el rango de valores posibles del parámetro μ es el intervalo (a_0, a_1) , la distribución inicial de referencia debe pues ser de la forma $\pi(\mu) = C\sqrt{n}/\sigma$ con C elegido de forma que $\int \pi(\mu) d\mu = 1$. Puesto que σ es una constante conocida, independiente de μ resulta, $C = \sigma/\{(a_1 - a_0)\sqrt{n}\}$ y $\pi(\mu) = 1/(a_1 - a_0)$, i.e. la distribución uniforme sobre el conjunto de sus posibles valores.

Consecuentemente, si se han realizado n observaciones independientes $\{x_1, \dots, x_n\}$ de una distribución normal $N(x|\mu, \sigma)$ con σ conocido, la distribución final de referencia será, haciendo uso del Teorema de Bayes con $\pi(\mu) = 1/(a_1 - a_0)$ como distribución inicial, y recordando la Ecuación 5.4.13,

$$\begin{aligned} \pi(\mu|\bar{x}) &\propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right\} \quad \text{para } \mu \in (a_0, a_1) \\ &= 0, \quad \text{fuera de ese intervalo.} \end{aligned} \quad (2)$$

Comparando con la Definición 4.3.5 de una distribución Normal, la distribución final de referencia resulta ser proporcional a la densidad normal $N(\mu|\bar{x}, \sigma/\sqrt{n})$ en el intervalo (a_0, a_1) de valores posibles de μ .

Debido al truncamiento producido en los límites del intervalo (a_0, a_1) deberemos obtener específicamente la correspondiente constante de proporcionalidad a partir de la ecuación

$$\int \pi(\mu|\bar{x}) d\mu = C \int_{a_0}^{a_1} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right\} d\mu = 1$$

En consecuencia,

$$\begin{aligned} C &= \left[\int_{a_0}^{a_1} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right\} d\mu \right]^{-1} = \\ &= \left[\frac{\sigma}{\sqrt{n}} \sqrt{2\pi} \int_{a_0}^{a_1} N(\mu|\bar{x}, \sigma/\sqrt{n}) d\mu \right]^{-1} = \\ &= \left\{ \frac{n}{2\pi\sigma^2} \right\}^{1/2} \left\{ \Phi\left(\frac{a_1 - \bar{x}}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a_0 - \bar{x}}{\sigma/\sqrt{n}}\right) \right\} \end{aligned}$$

de forma que introduciendo en (2) el valor de la constante hallada, y comparando con la definición de una densidad normal, la distribución final de referencia resulta ser

$$\pi(\mu|\bar{x}) = \frac{N(\mu|\bar{x}, \sigma/\sqrt{n})}{\Phi\left(\frac{a_1 - \bar{x}}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a_0 - \bar{x}}{\sigma/\sqrt{n}}\right)} \quad \text{para } \mu \in (a_0, a_1)$$

$$= 0, \quad \text{fuera de este intervalo.}$$

Si, en particular, no se conocen, o no quieren utilizarse, los límites para los valores posibles de μ , de forma que $a_0 \rightarrow -\infty$ y $a_1 \rightarrow +\infty$ el resultado anterior se reduce a $\pi(\mu|\bar{x}) = N(\mu|\bar{x}, \sigma/\sqrt{n})$.

Ejemplo 6.5.3. *Composición de la orina*

Determinar la probabilidad de que el contenido μ de creatina contenida en la orina de una mujer adulta sea superior a los 10 mg/24 h, basándose en los resultados experimentales 5, 15, y 19 mg/24 h que se suponen normalmente distribuidos con desviación típica 10 mg/24 h.

Haciendo uso de la distribución final de referencia que acaba de ser obtenida y del hecho obvio de que $\mu > 0$ y por tanto $(a_0, a_1) = (0, \infty)$, y teniendo en cuenta que en nuestro caso $\sigma/\sqrt{n} = 10/\sqrt{3} = 5,77$ y $\bar{x} = 13$ tendremos

$$p[\mu > 10] = \int_{10}^{\infty} \pi(\mu|\bar{x}) = \frac{\int_0^{\infty} N(\mu|13, 5,77) d\mu}{1 - \Phi(-13/5,77)} =$$

$$= \left\{ \Phi\left(\frac{10-13}{5,77}\right) - \Phi\left(\frac{-13}{5,77}\right) \right\} / \left\{ 1 - \Phi\left(\frac{-13}{5,77}\right) \right\} = 0,293$$

Como estudiamos en la Sección 5.5, la distribución de los resultados experimentales z depende frecuentemente de un parámetro *marginal desconocido* ω además de hacerlo del parámetro de interés θ , de forma que su distribución de probabilidad debe escribirse en la forma $p(z|\theta, \omega)$.

Si, en esta situación, deseamos obtener una distribución final de referencia para el parámetro de interés θ debemos determinar una distribución inicial de referencia $\pi(\theta, \omega)$ para *ambos* parámetros desconocidos, puesto que, de otra forma, el parámetro marginal ω no podría ser eliminado. Para construir esta distribución de referencia conjunta se procede de forma secuencial. En primer lugar, condicionando al parámetro de interés θ como si este fuese conocido, se determina la distribución condicional de referencia $\pi(\omega|\theta)$ del parámetro marginal; este resultado puede ser entonces usado para eliminar por integración el parámetro marginal obteniéndose así el modelo

$$p(z|\theta) = \int p(z|\theta, \omega) \pi(\omega|\theta) d\omega \quad (3)$$

que sólo depende de θ y al que pueden ser aplicados, por tanto, los procedimientos descritos en la sección anterior para obtener $\pi(\theta)$. La distribución inicial de referencia será entonces $\pi(\theta) \pi(\omega|\theta)$.

Naturalmente si, aunque no se disponga de información sobre θ , se dispone de información inicial sobre ω que permite asignar directamente una distribución inicial condicional $p(\omega|\theta)$, esta distribución puede ser utilizada en (3) en lugar de $\pi(\omega|\theta)$ para eliminar el parámetro marginal. En este caso, la distribución inicial de referencia será, obviamente, $\pi(\theta) p(\omega|\theta)$.

Puede demostrarse, por ejemplo, que si se utiliza este procedimiento con objeto de obtener la distribución final de referencia para la media μ de un modelo normal $N(\mu|\mu, \sigma)$ cuando la desviación típica σ es desconocida se obtiene la distribución condicional de referencia

$$\pi(\sigma|\mu) = \frac{1}{\log(b_1/b_0)} \frac{1}{\sigma} \quad \sigma \in (b_0, b_1)$$

$$= 0, \quad \text{fuera de ese intervalo}$$

donde (b_0, b_1) es el conjunto de valores posibles de σ ; la distribución inicial de referencia para μ resulta entonces ser, como en el caso en que σ era conocido,

$$\pi(\mu) = 1/(a_1 - a_0), \quad \mu \in (a_0, a_1)$$

$$= 0, \quad \text{fuera de ese intervalo}$$

donde (a_0, a_1) es el conjunto de valores posibles de μ . La correspondiente distribución final de referencia para μ cuando no se limitan los valores de σ , esto es cuando $b_0 \rightarrow 0$ y $b_1 \rightarrow \infty$, es de la forma

$$\pi(\mu|\bar{x}, s) = \frac{St(\mu|\bar{x}, s/\sqrt{(n-1)}, n-1)}{\Phi_{n-1}\left(\frac{a_1 - \bar{x}}{s/\sqrt{(n-1)}}\right) - \Phi_{n-1}\left(\frac{a_0 - \bar{x}}{s/\sqrt{(n-1)}}\right)}$$

donde Φ_{n-1} es la función de distribución de la distribución de Student $St(\mu|0, 1, n-1)$ que está extensivamente tabulada. Cuando, en particular, $a_0 \rightarrow -\infty$ y $a_1 \rightarrow +\infty$ reaparece la distribución de referencia

$$\pi(\mu|\bar{x}, s) = St(\mu|\bar{x}, s/\sqrt{(n-1)}, n-1)$$

que ya obtuvimos en la Sección 5.5.

En el Teorema 6.5.4 se recogen los resultados anteriores y los correspondientes a otros modelos probabilísticos frecuentes, cuya demostración dejamos como ejercicio.

TEOREMA 6.5.4. *Distribuciones finales de referencia*

Modelo probabilístico	Estadístico suficiente	Distrib. inicial de referencia	Distrib. final de referencia
Bernoulli $Br(x \theta)$	(r, n) $r = \sum x_i$	$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$	$Be(\theta r+1/2, n-r+1/2)$ $0 < \theta < 1$
Poisson $Po(x \lambda)$	(r, n) $r = \sum x_i$	$\pi(\lambda) \propto \lambda^{-1/2}$	$Ga(\lambda r+1/2, n)$ $\lambda > 0$
Normal $N(x \mu, \sigma)$ σ conocido	(\bar{x}, n) $\bar{x} = \sum x_i/n$	$\pi(\mu) \propto 1$	$N(\mu \bar{x}, \sigma/\sqrt{n})$ $-\infty < \mu < \infty$
Normal $N(x \mu, \sigma)$ μ conocido	(s, n) $s = \sum(x_i - \mu)^2/n$ $b = 1/\sigma^2$	$\pi(b) \propto b^{-1}$	$Ga(b n/2, ns^2/2)$ $b > 0$
Normal $N(x \mu, \sigma)$	(s, \bar{x}, n) $s = \sum(x_i - \bar{x})^2/n$	$\pi(\mu, b) \propto b^{-1}$	$St(\mu \bar{x}, s/\sqrt{(n-1)}, n-1)$ $-\infty < \mu < \infty$
Normal $N(x \mu, \sigma)$	(s, n) $s = \sum(x_i - \bar{x})^2/n$	$\pi(\mu, b) \propto b^{-1}$	$Ga(b (n-1)/2, ns^2/2)$ $b > 0$
Exponencial $Ex(x \theta)$	(r, n) $r = \sum x_i$	$\pi(\theta) \propto \theta^{-1}$	$Ca(\theta r, r)$ $\theta > 0$

Estas distribuciones de referencia se refieren al caso en que no se imponen límites al campo de variabilidad del parámetro de interés distintos de los exigidos por el propio modelo probabilístico; si se limita dicho campo se obtienen distribuciones de referencia proporcionales a las anteriores, con la constante de proporcionalidad ajustada de forma que continúen integrando la unidad.

6.6. Análisis de sensibilidad

A lo largo de nuestro estudio de los métodos Bayesianos de inferencia hemos procurado subrayar el hecho de que el resultado final, esto es la distribución final del parámetro de interés y por tanto la decisión óptima, dependen de la distribución inicial; este hecho, que algunos autores han descrito como un defecto de la metodología Bayesiana constituye en realidad uno de sus mayores atractivos.

En efecto, esta dependencia es precisamente la que permite incorporar la información inicial que frecuentemente posee el investigador. Esta informa-

ción es muchas veces crucial; piénsese por ejemplo que en el diseño de un experimento o en las decisiones relativas a una enfermedad rara, no suele disponerse de ningún otro tipo de información. Sin embargo, la importancia relativa de esta información inicial con respecto a la proporcionada por los datos puede y debe ser medida para poder apreciar sus consecuencias y, en particular, para poder distinguir entre las conclusiones que se basan en los resultados experimentales que se analizan y las que dependen de la experiencia acumulada por el investigador. Como podía esperarse, las distribuciones finales de referencia descritas en la sección anterior constituyen la herramienta natural para abordar este problema.

En efecto, representando en una misma escala la distribución inicial $p(\theta)$, que describe la información de que dispone el investigador sobre el valor del parámetro de interés θ , su correspondiente distribución final $p(\theta|z)$ y la distribución final de referencia $\pi(\theta|z)$, puede observarse inmediatamente la importancia relativa que tiene en la práctica la información inicial; cuanto más alejada esté $p(\theta|z)$ de la distribución de referencia $\pi(\theta|z)$ (que, recordemos, describe únicamente la información proporcionada por los datos) más importancia práctica tendrá la información inicial del investigador y más cuidado, por tanto, deberá tenerse al tratar de cuantificarla.

La discrepancia mencionada puede ser medida haciendo uso de la Definición 6.3.1. Así,

$$D\{p(\theta|z), \pi(\theta|z)\} = \int p(\theta|z) \log \frac{p(\theta|z)}{\pi(\theta|z)} d\theta$$

es una descripción cuantitativa de la importancia práctica de la información inicial contenida en la distribución $p(\theta)$ cuando se analizan con ella los resultados experimentales z .

El lector de un trabajo en el que se describen los resultados de una investigación debe estar en condiciones de poder analizar la sensibilidad de las conclusiones a cambios en la información inicial. En particular, debe poder distinguir entre las conclusiones que se derivan de los resultados experimentales que se le comunican y las que dependen de la experiencia previa de los autores de la investigación. Con este objeto, un trabajo científico que base sus conclusiones en el análisis de un conjunto de datos experimentales debe contener, al menos, los siguientes elementos.

- a) Descripción detallada del parámetro de interés θ , del experimento e que ha sido realizado para obtener información sobre θ , y de los datos obtenidos, z .
- b) Descripción del modelo probabilístico utilizado para su análisis $p(z|\theta)$ y de las razones en que se basa su adopción.

- c) Valor numérico del estadístico suficiente $t = t(z)$ que resume, si se acepta el modelo propuesto, toda la información sobre θ contenida en los datos z .
- d) Distribución inicial $p(\theta)$ que describe la información inicial del equipo investigador sobre el valor de θ .
- e) Distribución final de referencia $\pi(\theta|z)$ que describe la información sobre θ proporcionada por los datos si el modelo es correcto.
- f) Distribución final $p(\theta|z)$ que combina la información inicial sobre θ con la que proporcionan los datos.
- g) Conclusiones numéricas sobre la importancia relativa de la información inicial que se derivan de la discrepancia entre $p(\theta|z)$ y $\pi(\theta|z)$.

Con esta información, el lector dispone de varias opciones que le permiten en todo caso extraer sus conclusiones a la vista de los resultados experimentales descritos.

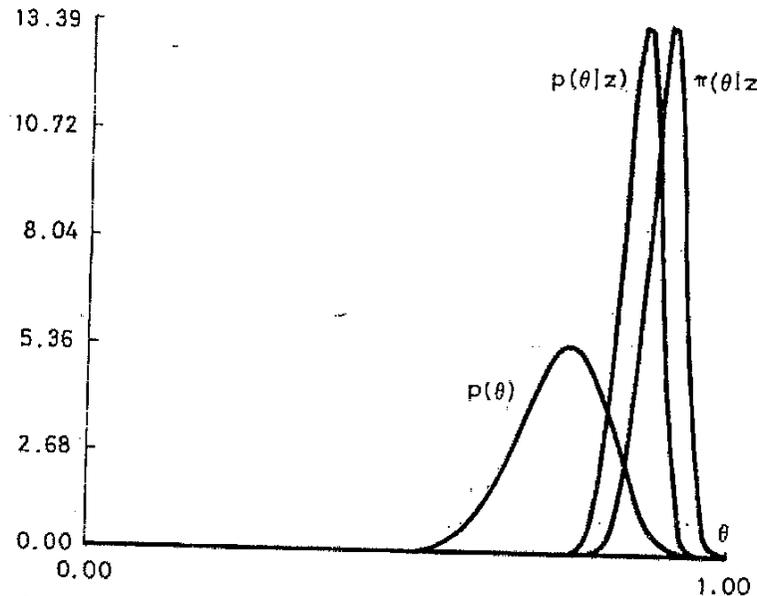
- a) Si la elección del modelo probabilístico no le parece correcta, debe partir directamente de los datos originales z y analizarlos por sí mismo siguiendo el proceso descrito.
- b) Si acepta el modelo pero no hace suya la información inicial del equipo que ha realizado el trabajo, puede basar sus conclusiones en la distribución final de referencia $\pi(\theta|z)$, o bien especificar su propia distribución inicial, calcular por su cuenta la correspondiente distribución final, y basar sus decisiones en ella.
- c) Finalmente, si acepta como válida la información experimental y hace suya la información inicial del equipo que ha realizado el trabajo, puede basar sus decisiones en la distribución final $p(\theta|z)$ que tal equipo ha obtenido. En este caso, la discrepancia entre $p(\theta|z)$ y $\pi(\theta|z)$ le proporciona una medida del nivel de riesgo a que se somete aceptando como válida tal información inicial.

Ejemplo 6.6.1. Resultados de un muestreo

Con objeto de determinar la proporción θ de la población del área metropolitana de la ciudad a los que les ha sido aplicada alguna vez una determinada vacuna se procede a un muestreo aleatorio de la población. La información inicial del equipo responsable de la investigación puede ser descrita mediante la distribución inicial $Be(\theta|25, 9)$ y de 100 personas encuestadas 90 reconocen haber sido vacunadas. Describir adecuadamente los resultados de la investigación.

Si las 100 personas encuestadas han sido escogidas de forma aleatoria, esto es de forma que inicialmente todos los elementos de la población tenían la misma probabilidad de ser escogidos en la muestra, se trata de 100 observaciones de una distribución de Bernoulli con parámetro θ . En consecuencia, si r es el número de personas vacunadas en la muestra, la distribución final de referencia es (Sección 6.5) $Be(\theta|r + 1/2, n - r + 1/2)$

DATOS VACUNACION



	Inicial, $p(\theta)$.	Referencia $\pi(\theta z)$	Final, $p(\theta z)$
Parámetros	25, 9	90,5, 10,5	115, 19
Moda	0,7500	0,9040	0,8636
Media	0,7353	0,8960	0,8582
Mediana	0,7400	0,8987	0,8600
$I_{0,5}(\theta)$	0,697, 0,798	0,883, 0,923	0,843, 0,883
$I_{0,95}(\theta)$	0,615, 0,858	0,848, 0,945	0,810, 0,908
$I_{0,99}(\theta)$	0,588, 0,875	0,836, 0,952	0,799, 0,915
$I_{0,999}(\theta)$	0,533, 0,904	0,811, 0,963	0,776, 0,928

v. si la distribución es $Be(\theta|\alpha, \beta)$, la correspondiente distribución final es (Sección 6.2) $Be(\theta|\alpha + r, \beta + n - r)$. En nuestro caso tenemos $n = 100$, $r = 90$, $\alpha = 25$ y $\beta = 9$; en consecuencia,

$$\pi(\theta|r) = Be(\theta|90, 5, 10, 5)$$

$$p(\theta) = Be(\theta|25, 9)$$

$$p(\theta|r) = Be(\theta|115, 19)$$

Una descripción adecuada de los resultados de esta investigación debe incluir, al menos, el modelo supuesto, el valor del estadístico suficiente, la representación gráfica de las tres densidades de probabilidad encontradas y los valores numéricos de sus características más notables. En la figura de la página anterior se recogen estos elementos en la forma en que aparecen como resultado de un programa de ordenador que realiza automáticamente el análisis cuando el modelo pertenece a la clase exponencial y la distribución inicial pertenece a la familia conjugada correspondiente.

Estos resultados, que pueden ser calculados a mano sin dificultad alguna con ayuda de unas tablas, deberían ir naturalmente acompañados de un informe, redactado por el equipo responsable de la investigación, que detalle las implicaciones sociológicas o epidemiológicas que se derivan de los resultados numéricos obtenidos. En el ejemplo que nos ocupan la moda de la distribución final de referencia es 0,9040 mientras que la de la distribución final es 0,8636. En consecuencia, en el caso en que sea importante discriminar entre los valores de θ comprendidos entre estos límites deberá procederse a analizar críticamente el contenido de la información inicial pues la decisión final, en este caso, podría depender crucialmente de ella.

Concluiremos esta sección subrayando el hecho obvio de que el análisis de la sensibilidad de la distribución final del parámetro de interés a cambios en su distribución inicial, o en el modelo probabilístico utilizado, debe hacerse teniendo en cuenta el uso que va a hacerse de las conclusiones obtenidas. En efecto, como hemos mencionado repetidamente, los problemas de inferencia son generalmente pasos intermedios (necesarios para incorporar nueva información experimental) en un proceso de decisión y, obviamente, la importancia de pequeñas variaciones en $p(\theta|z)$ dependerá de las variaciones que introduzca en las utilidades esperadas de las distintas decisiones bajo consideración.

6.7. Discusión y referencias

Debido a la complejidad matemática de las distribuciones finales que aparecen como resultado de los análisis de datos experimentales, resulta casi imprescindible recurrir a sus representaciones gráficas y al cálculo de algunas de sus características para poder comunicar su contenido más relevante; el carácter intuitivo de las regiones creíbles hace que estas características sean frecuentemente las más apropiadas cuando se trata de transmitir los rasgos esenciales de las conclusiones alcanzadas.

Por otra parte, hemos observado que el análisis de las distribuciones finales resulta más sencillo si el modelo probabilístico pertenece a la clase ex-

ponencial; Darms (1935), Koopman (1936) y Pitman (1936) demostraron que entonces, y bajo ciertas condiciones de regularidad sólo entonces, existe un estadístico suficiente cuya dimensión no depende del tamaño de la muestra. El concepto de estadístico suficiente fue introducido por Fisher (1922) y estudiado en profundidad por Halmos & Savage (1949); aunque la definición clásica de suficiencia es distinta de la enunciada en la Sección 6.2, su motivación es la misma y puede demostrarse que, de hecho, se trata de dos definiciones equivalentes (Lindley, 1965, § 5.5).

Si el modelo probabilístico admite una familia conjugada de distribuciones (para lo que es suficiente, aunque no necesario, que el modelo pertenezca a la clase exponencial), resulta tentador describir la información inicial mediante un elemento de esta familia; esta operación, sin embargo, no debe nunca hacerse de forma mecánica: es fundamental comprobar que, efectivamente, la información inicial puede ser descrita, al menos aproximadamente, por un miembro de tal familia. El concepto de familia conjugada de distribuciones fue formalizado por Raiffa & Schlaifer (1961); en DeGroot (1970, cap. 9) se estudian con detalle numerosos ejemplos.

El estudio de las transformaciones del parámetro de interés cuya distribución puede aproximarse bien mediante una distribución normal tiene una larga, pero dispersa historia (ver e.g., Kendall & Stuart, 1958/1977); Bernardo (1980 a) aborda el problema desde la perspectiva de la teoría de la decisión.

El estudio del comportamiento asintótico de la distribución final cuando el tamaño n de la muestra tiende a infinito es importante por varias razones: a nivel práctico, proporciona aproximaciones útiles cuando el tamaño real de la muestra es suficientemente grande; permite además demostrar que la mayor parte de las recetas clásicas son aproximaciones razonables si, y solamente si, se dispone de muestras grandes; finalmente, su estudio es un requisito para la obtención de distribuciones de referencia. El comportamiento asintótico de las distribuciones finales ha sido estudiado por LeCam (1953, 1958, 1970), Lindley (1961), Anscombe (1964), Walker (1969), Johnson (1967, 1970) y Dawid (1970).

Se ha trabajado mucho en la obtención de distribuciones iniciales que añadan poca información a la información experimental, empezando por los trabajos de Bayes (1763) y Laplace (1812/1912) basados en consideraciones de simetría. Intentos más modernos se basan en exigencias de invarianza (Jeffreys, 1939/1967; Barnard, 1952; Hartigan, 1964; Box & Tiao, 1973, § 1.3 y Jaynes, 1980), en límites de distribuciones conjugadas (Novick, 1969; DeGroot, 1970, cap. 10) o en argumentos basados en la teoría de la información (Jaynes, 1968; Good, 1969; Zellner, 1977 y Bernardo, 1979 b). Estos trabajos han demostrado que la mayor parte de los resultados numéricos clásicos

encontrados en la práctica son casos particulares de la metodología Bayesiana. Así, por ejemplo, los estimadores y los intervalos de confianza clásicos son, muy frecuentemente, los valores medios y las regiones creíbles, respectivamente, de las correspondientes distribuciones finales de referencia (Lindley, 1965; Welch & Peers, 1963; Bartholomew, 1965).

El análisis de la sensibilidad de la distribución final a cambios en la distribución inicial debería formar parte, de manera sistemática, del contenido de todo análisis científico de datos experimentales. El análisis de la sensibilidad de la distribución final a cambios en el modelo probabilístico utilizado es más complejo y constituye una de las más importantes líneas de investigación actuales. Un enfoque prometedor es el de Smith & Spiegelhalter (1980).

PROBLEMAS

1. Se sabe que el tiempo de espera entre el registro de dos emisiones consecutivas de un trazador radioactivo es una cantidad aleatoria t con densidad de probabilidad $p(t|\theta) \propto e^{-\theta t}$, $t > 0,1$ y nula para $t < 0,1$ debido a que el contador permanece bloqueado una décima de segundo cuando registra una señal. Si la información inicial sobre θ puede describirse mediante la distribución $Ga(\theta|1, 1)$, determinar la región creíble para θ de nivel 0,95, tras observar registros en los tiempos 1,2, 2,3, 3,5 y 4,8.
2. Se sabe que el número x de microorganismos en una zona de una preparación microscópica de área S es una cantidad aleatoria con distribución $p(x|\theta) = Po(x|\theta S)$. Procediendo al conteo de 10 cuadrículas de 1 mm^2 se encuentra un valor medio de 3 microorganismos por cuadrícula. Si la distribución inicial sobre θ es $Ga(\theta|8, 2)$, determinar la probabilidad de que no exista ningún microorganismo en una zona de 3 mm^2 de superficie.
3. Determinar aproximadamente, haciendo uso de una transformación normalizadora una distribución Gamma para una cantidad aleatoria X cuyo valor más probable sea 4 y tal que $P\{X > 8\} = 0,1$.
4. Los datos obtenidos mediante una encuesta sobre el resultado de las próximas elecciones locales son de 200 personas aleatoriamente elegidas 82 votarán a la izquierda, 75 a la derecha y el resto se abstendrán. Si la distribución inicial sobre la proporción θ de votos para la izquierda entre los votos válidos puede describirse mediante una distribución $Be(\theta|60, 55)$, determinar la probabilidad de que la izquierda gane las elecciones.
5. La información inicial sobre la proporción θ de enfermos que sobreviven a un determinado tratamiento puede describirse mediante una distribución $Be(\theta|80, 2)$. Revisadas 1.850 historias clínicas de pacientes sometidas a ese tratamiento se encuentra que sobrevivieron el 96%. Determinar, aproximadamente, los intervalos creíbles para θ de niveles 0,9, 0,99 y 0,999.

6. Con objeto de determinar la desviación típica σ correspondiente a un determinado aparato de análisis espectrográfico, se realizan 2.000 observaciones $N(x|\mu, \sigma^2)$ de una preparación de la que se conoce el contenido exacto μ de la sustancia objeto de medida, encontrándose una desviación típica muestral $s = \sqrt{\{\sum(x_i - \mu)^2/n\}} = 2,2 \text{ mg}$. Utilizando la correspondiente distribución asintótica, determinar aproximadamente la probabilidad de que σ sea mayor que 2,25 mg.
7. Se duda entre θ_1 y θ_2 como causas del síndrome que presenta un determinado enfermo, pero se sabe que θ_1 es *al menos* dos veces más probable que θ_2 . Realizando un análisis, se obtienen unos resultados θ cuya función de verosimilitud viene dada por $p(x|\theta_1) = 0,8$, $p(x|\theta_2) = 0,6$. Determinar la distribución final de referencia para θ .
8. Se sabe que la cantidad θ de una determinada sustancia en la sangre no puede sobrepasar los 3 mg/litro. Analizadas 20 personas, se encuentra un valor medio de 2 mg. Suponiendo que se trata de observaciones de una población normal con desviación típica 0,8 mg, determinar la distribución final de referencia para θ .
9. Comparar, gráficamente y analíticamente, el resultado anterior con el que se obtendría utilizando una información inicial sobre θ descrita mediante la distribución

$$p(\theta) \propto N(\theta|1,8, 0,4), \quad 0 < \theta < 3$$

$$= 0 \text{ en caso contrario.}$$

10. La opinión inicial de un equipo de farmacólogos sobre la duración en días t de un determinado tipo de preparado es que se descompondrá al cabo de unos 50 días y que, con probabilidad 0,95, no llegará a los 70. Preparadas 10 muestras, se descomponen a los 40, 45, 52, 38, 60, 55, 42, 48, 56 y 49 días. Realizar un análisis completo de estos datos, suponiendo que se trata de observaciones normales, de desviación típica 14.

Análisis cuantitativo de decisiones

En este capítulo, los conceptos desarrollados hasta ahora son utilizados para analizar determinados problemas específicos de decisión.

Se estudian los problemas clásicos de *contraste de hipótesis* y de *estimación puntual* y se demuestra que la inferencia estadística puede ser considerada como un problema de decisión cuya función de utilidad es una medida de información.

Se describen métodos para especificar la función de utilidad, se determina el *valor de la información* proporcionada por los datos experimentales, y se aborda el problema del *diseño de experimentos*.

Se analizan finalmente las características fundamentales de los problemas de decisión más frecuentes en la práctica médica.

En el capítulo de Fundamentos fue demostrado que el único procedimiento de tomar decisiones que es consistente con los principios de coherencia exige cuantificar la información sobre los sucesos inciertos con una distribución de probabilidad, precisar las preferencias entre las consecuencias posibles mediante una función de utilidad, y elegir aquella decisión que maximiza la utilidad esperada; en los capítulos restantes ha sido estudiada la forma de describir, mediante una distribución de probabilidad, la información relevante de que se dispone incluyendo, en su caso, la información experimental. En este último capítulo estudiaremos la estructura de determinados problemas concretos de decisión y analizaremos sus soluciones.

7.1. Contraste de hipótesis y estimación puntual

Un tipo de problema de decisión muy frecuente, y extraordinariamente sencillo, es el que se presenta cuando debe elegirse entre dos modelos o hi-

pótesis alternativas H_0 y H_1 . En este caso, tanto el espacio de decisiones D como el espacio de sucesos inciertos Θ contienen únicamente éstos dos elementos, esto es $D = \Theta = \{H_0, H_1\}$, y la función de utilidad es de la forma indicada en la tabla.

	Θ	
D	H_0	H_1
H_0	u_{00}	u_{01}
H_1	u_{10}	u_{11}

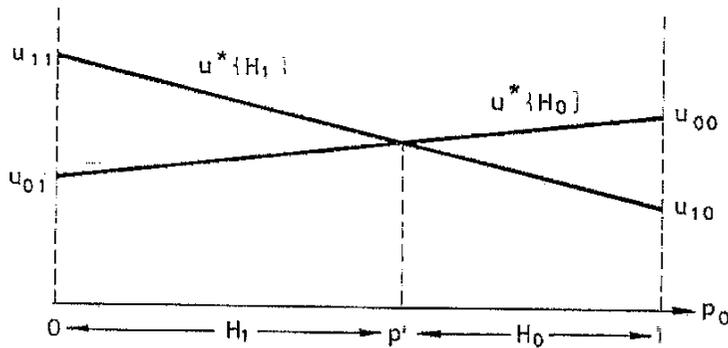
Resulta así, por ejemplo, que la utilidad de aceptar la hipótesis H_0 cuando es cierta es u_{00} , mientras que la utilidad de rechazar H_0 (esto es aceptar H_1) cuando H_0 es cierta es u_{01} . Esta es la estructura básica del tipo de problemas de decisión generalmente conocidos como problemas de *contraste de hipótesis*.

Si llamamos p_0 a la probabilidad asociada a la hipótesis H_0 en el momento de tomar la decisión, la utilidad esperada de aceptar cada una de las dos hipótesis vendrá dada por

$$u^*\{H_0\} = u_{00} p_0 + u_{01}(1 - p_0)$$

$$u^*\{H_1\} = u_{10} p_0 + u_{11}(1 - p_0)$$

cuya representación gráfica es del tipo



En consecuencia, la decisión óptima es aceptar la hipótesis H_0 si, y solamente si

$$u_{00} p_0 + u_{01}(1 - p_0) > u_{10} p_0 + u_{11}(1 - p_0)$$

esto es si la probabilidad p_0 asociada a la hipótesis H_0 en el momento de tomar la decisión es mayor que p' , donde

$$p' = \frac{u_{11} - u_{01}}{u_{00} + u_{11} - u_{01} - u_{10}}$$

Si $p_0 = p'$ las dos decisiones son equivalentes y si $p_0 < p'$ la decisión óptima es H_1 .

En muchas ocasiones, $u_{00} = u_{11} = u$; si definimos entonces $\alpha = u - u_{01}$ y $\beta = u - u_{10}$, de forma que α es la *pérdida de oportunidad* en que se incurre si se acepta H_0 cuando es falsa y β la pérdida de oportunidad en que se incurre si se rechaza H_0 cuando es cierta, la decisión óptima resulta ser H_0 si, y solamente si,

$$\frac{p_0}{1 - p_0} > \frac{\alpha}{\beta}$$

una condición cuyo contenido intuitivo es aparente.

Ejemplo 7.1.1. Comercialización de un fármaco

En un laboratorio farmacéutico quiere contrastarse la hipótesis, H_0 , de que la proporción θ de veces en las que un nuevo preparado resulta eficaz frente a una determinada enfermedad es mayor que 0,95. Si se acepta H_0 siendo cierta, se obtiene un beneficio de 10 millones debido a la comercialización de un producto útil, mientras que si se acepta siendo falsa, se incurre en una pérdida de 7 millones debida a que el producto, ya comercializado, debe ser retirado del mercado. Si se rechaza H_0 el producto no es comercializado y se pierden por tanto las 500.000 pesetas invertidas en la investigación. La información obtenida extrapolando resultados experimentales realizados sobre animales puede describirse mediante la distribución $Be(\theta|15, 3)$. Probado el nuevo preparado sobre 50 pacientes, se observa una reacción positiva en 47 de ellos. Determinar la decisión óptima, suponiendo la utilidad proporcional al dinero.

La distribución final de θ es, de acuerdo con el Teorema 6.2.2,

$$p(\theta|z) = Be(\theta|\alpha + r, \beta + n - r) = Be(\theta|62, 6)$$

y en consecuencia

$$p_0 = p(H_0) = \int_{0.95}^1 Be(\theta|62, 6) d\theta \approx 0,106$$

como puede comprobarse utilizando el Teorema 6.3.2. La función de utilidad es de la forma

$$u\{H_0|H_0\} = u_{00} = 10, \quad u\{H_0|H_1\} = u_{01} = -7$$

$$u\{H_1|H_0\} = u_{10} = u\{H_1|H_1\} = u_{11} = -0,5$$

en consecuencia el valor mínimo p de la probabilidad de la hipótesis H_0 para aceptarla debe ser

$$(u_{11} - u_{01}) / (u_{00} + u_{11} - u_{10} - u_{01}) = 6,5/17 = 0,382$$

que es mayor que p_0 ; en consecuencia, la hipótesis H_0 debe ser rechazada y el nuevo fármaco no comercializado.

El problema de decisión que acabamos de describir es un caso particular de una amplia clase de problemas de decisión en los que coinciden el espacio de decisiones y el espacio de sucesos inciertos. Esta situación es especialmente frecuente en la investigación científica; en efecto, el interés primordial del investigador se centra frecuentemente en determinar el valor verdadero de una magnitud desconocida θ , de forma que el espacio de decisión D coincide con el conjunto Θ de valores posibles de tal magnitud. Los problemas de decisión en los que coinciden los espacios D y Θ son generalmente denominados problemas de *estimación puntual* debido al hecho de que la decisión final consiste en elegir un punto $\hat{\theta} \in \Theta$ que constituye una estimación del verdadero, y desconocido, valor de θ .

Para completar la descripción del problema de decisión es necesario especificar una distribución de probabilidad que describa la información del decisor sobre el parámetro desconocido θ en el momento de tomar la decisión, y una función de utilidad que especifique las preferencias del decisor entre las posibles consecuencias de su decisión.

Una de las funciones de utilidad empleadas con más frecuencia en esta clase de problemas es la de tipo *cuadrático*, en la que se supone que la utilidad $u(\hat{\theta}, \theta)$ de estimar mediante $\hat{\theta}$ el valor verdadero θ del parámetro de interés es de la forma

$$u(\hat{\theta}, \theta) = B(\theta) - A(\hat{\theta} - \theta)^2, \quad A > 0 \quad (1)$$

donde tanto la constante A como la función $B(\theta)$ son arbitrarias. De esta forma la *pérdida* ocasionada por el error de estimación crece proporcionalmente al *cuadrado* del error cometido.

En esta situación, si $p(\theta|H)$ es la función que describe la información de que se dispone sobre el valor de θ en las condiciones H en que debe tomarse la decisión (*), la utilidad esperada de tomar la decisión $\hat{\theta}$ vendrá dada por

(*) Obviamente, $p(\theta|H)$ será una función de probabilidad si θ es una cantidad aleatoria discreta y una función de densidad de probabilidad si se trata de una cantidad aleatoria continua.

$$u^*(\hat{\theta}) = \sum \{ B(\theta_i) - A(\hat{\theta} - \theta_i)^2 \} p(\theta_i|H) \quad (2)$$

si θ es una cantidad aleatoria discreta y por

$$u^*(\hat{\theta}) = \int \{ B(\theta) - A(\hat{\theta} - \theta)^2 \} p(\theta|H) d\theta \quad (3)$$

si θ es continua.

TEOREMA 7.1.1. *La mejor estimación puntual de θ con pérdida cuadrática es la media de la distribución de θ en el momento de producirse la estimación.*

Demostración

Nos limitaremos al caso discreto; la demostración para el caso continuo es totalmente análoga. Debemos encontrar el máximo de (2), esto es, el valor de $\hat{\theta}$ que maximiza

$$\sum B(\theta_i) p(\theta_i|H) - A \sum (\hat{\theta} - \theta_i)^2 p(\theta_i|H)$$

Iguando a cero la derivada respecto de $\hat{\theta}$ obtenemos la ecuación

$$-2A \sum (\hat{\theta} - \theta_i) p(\theta_i|H) = 0$$

esto es

$$\hat{\theta} \sum p(\theta_i|H) = \sum \theta_i p(\theta_i|H)$$

de forma que, para ser un extremo de (2), el valor de $\hat{\theta}$ debe coincidir con la media de la distribución de θ . Puede comprobarse además, derivando de nuevo, que se trata efectivamente de un máximo.

Naturalmente, (1) no es la única función de utilidad que puede ser considerada. Otra función de utilidad frecuentemente utilizada exige que la *pérdida* ocasionada por el error de estimación sea proporcional al *valor absoluto* del error, de forma que

$$u\{\hat{\theta}, \theta\} = B(\theta) - A|\hat{\theta} - \theta|, \quad A > 0$$

Es posible entonces encontrar un resultado análogo al Teorema 7.2.1, cuya demostración omitiremos.

TEOREMA 7.1.2. *La mejor estimación puntual de θ con pérdida absoluta es la mediana de la distribución de θ en el momento de producirse la estimación.*

Obviamente, si $p(\theta|H)$ es una distribución simétrica, su media y su mediana coinciden y encontramos en ambos casos el mismo estimador.

Ejemplo 7.1.2. Control de calidad

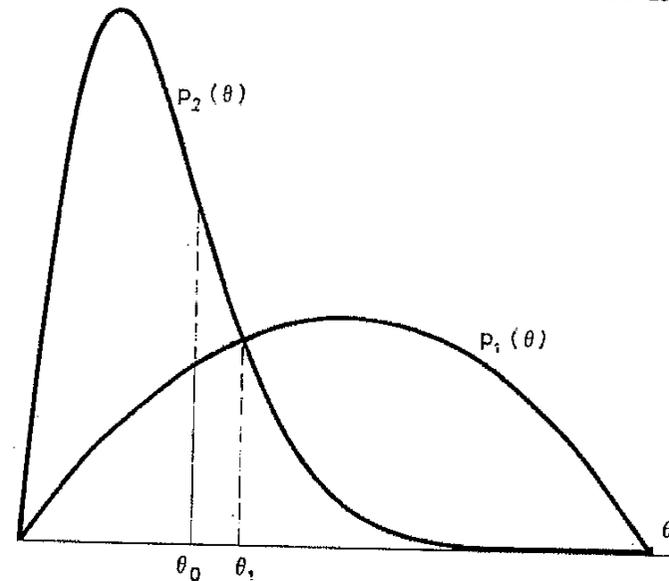
En una cadena de producción quiere estimarse la proporción θ de elementos defectuosos producidos. Suponiendo que las pérdidas en que se incurre como consecuencia de una estimación errónea son proporcionales al cuadrado del error cometido, determinar la estimación óptima de θ si se han observado 5 elementos defectuosos en una muestra de 70 y no se dispone de información adicional alguna.

Se trata de un proceso de Bernoulli con parámetro θ del que se ha observado una muestra de 70 elementos con $r = \sum x_i = 5$. En consecuencia (Teorema 6.5.4) la distribución final de referencia, que describe la información sobre θ de que se dispone en el momento de decir, es $Be(\theta|5 + 0,5, 70 - 5 + 0,5) = Be(\theta|5,5, 65,5)$ cuyo valor esperado es 0,0775. La producción de elementos defectuosos debe pues ser estimada en un 7,75 %.

7.2. La inferencia estadística como problema de decisión

En la sección anterior hemos descrito la situación en que el decisor está interesado en una *estimación* del verdadero valor del parámetro de interés θ . Frecuentemente sin embargo, especialmente en problemas de investigación y decisor está interesado en obtener cuanta información le sea posible sobre el parámetro de interés, de forma que no resulta razonable limitarse a una estimación puntual de su valor. Como ya mencionamos en la Sección 6.6, el equipo que culmina una etapa de investigación debe comunicar sus conclusiones finales sobre la magnitud de interés, junto con la argumentación en que tales conclusiones descansan. En términos más técnicos, los investigadores deben comunicar su distribución final para θ , junto con el modelo, los datos y la distribución inicial utilizados para obtenerla. Por parte de la comunidad científica, la utilidad de sus conclusiones vendrá medida por la cantidad de nueva información proporcionada. En consecuencia, *el problema de inferencia sobre θ puede ser descrito como un problema de decisión, cuyo espacio paramétrico es el conjunto Θ de valores posibles de θ , cuyo espacio de acciones es la familia de las posibles distribuciones finales de θ , y cuya función de utilidad es una cierta medida de información.*

Para precisar la estructura del problema de decisión descrito es necesario especificar la forma concreta de la función $u\{\hat{p}(\theta), \theta_0\}$ que describe la utilidad de describir mediante $\hat{p}(\theta)$ la información obtenida sobre la magnitud de interés cuando θ_0 es su verdadero valor.



Es obvio, por ejemplo, que una situación como la representada en la figura la utilidad de $p_2(\theta)$ debe ser mayor que la de $p_1(\theta)$ cuando θ_0 sea el verdadero valor de θ .

Resulta asimismo natural, en un modelo que pretende describir el proceso de investigación científica, recurrir a funciones de utilidad que fomenten la honestidad del científico. Puesto que la utilidad esperada por el investigador si comunica sus resultados mediante la distribución $p(\theta)$, cuando su información sobre θ en las condiciones H en que lo hace está descrita por la distribución $p(\theta|H)$, viene dada por

$$\sum u\{\hat{p}(\theta_i), \theta\}p(\theta_i|H), \text{ si } \theta \text{ es discreta}$$

$$\int u\{\hat{p}(\theta), \theta\}p(\theta|H)d\theta, \text{ si } \theta \text{ es continua}$$

y, puesto que la única postura coherente del investigador es maximizar su utilidad esperada, resulta obvio exigir que la función de utilidad u satisfaga la ecuación

$$\sup_{\hat{p}(\theta)} \sum u\{\hat{p}(\theta_i), \theta\}p(\theta_i|H) = \sum u\{p(\theta_i|H), \theta\}p(\theta_i|H), \text{ si } \theta \text{ es discreta} \quad (1a)$$

$$\sup_{\hat{p}(\theta)} \int u\{\hat{p}(\theta), \theta\}p(\theta|H)d\theta = \int u\{p(\theta|H), \theta\}p(\theta|H)d\theta, \text{ si } \theta \text{ es continua} \quad (1b)$$

de forma que la utilidad esperada del científico sea maximizada cuando $\hat{p}(\theta) = p(\theta|E)$, esto es cuando comunica sus verdaderas opiniones. Se dice que una función de utilidad es *propia* cuando verifica la ecuación (1a) o (1b):

Se conocen muchas funciones propias de utilidad; las expresiones siguientes para $u\{p(\theta), \theta_0\}$ definen algunas de ellas.

$$A \log\{p(\theta_0)\} + B(\theta_0) \quad (\text{logarítmica})$$

$$A\{2p(\theta_0) - |p(\theta)|_2^2\} + B(\theta_0) \quad (\text{cuadrática})$$

$$\frac{A}{\alpha - 1} \left[\left\{ \frac{p(\theta_0)}{|p(\theta)|_\alpha} \right\}^{\alpha-1} - 1 \right] + B(\theta_0) \quad (\text{esférica, } \alpha > 1)$$

donde $|p(\theta)|_\alpha = \{\sum p^\alpha(\theta_i)\}^{1/\alpha}$ si θ es discreta y $|p(\theta)|_\alpha = \{\int p^\alpha(\theta)d\theta\}^{1/\alpha}$ si es continua, y donde la constante A y la función $B(\theta)$ son arbitrarias.

La adopción de una de estas funciones, o de cualquier otra función de utilidad propia, dependerá de las razones que motiven el problema de inferencia. La función de utilidad esférica, por ejemplo, fue utilizada en la Sección 3.6 para obtener un método razonable de calificar exámenes constituidos por cuestiones con respuesta múltiple.

Escogiendo la función $B(\theta)$ de forma que sea una función adecuada de la (densidad de) probabilidad asociada al verdadero valor del parámetro de interés por una distribución origen $p_0(\theta)$, esto es de forma que $B(\theta) = f\{p_0(\theta)\}$, puede conseguirse que las funciones de utilidad sean invariantes frente a transformaciones monótonas de θ y que satisfagan, además, la condición normalizadora $u\{p_0(\theta), \theta\} = 0$ para todo θ de forma que, dado θ , se obtienen utilidades positivas si se consiguen distribuciones $p(\theta)$ «mejores» que la distribución $p_0(\theta)$ tomada como origen. Las tres familias de funciones de utilidad propias antes mencionadas se convierten entonces en

$$A \log\{p(\theta_0)/p_0(\theta_0)\} \quad (\text{logarítmica})$$

$$A \left\{ 2 \frac{p(\theta_0)}{p_0(\theta_0)} - \int \frac{p^2(\theta)}{p_0(\theta)} d\theta - 1 \right\} \quad (\text{cuadrática})$$

$$\frac{A}{\alpha - 1} \left[\left\{ \frac{p(\theta_0)/p_0(\theta_0)}{\int \{p(\theta)/p_0(\theta)\}^\alpha p(\theta)d\theta} \right\}^{\alpha-1} - 1 \right] \quad (\text{esférica, } \alpha > 1)$$

Puede demostrarse además que la función logarítmica es el límite de las esféricas cuando $\alpha \rightarrow 1$.

Consideremos, finalmente, una situación de inferencia «pura», esto es una situación en la que el científico está interesado en obtener información sobre θ sin pensar en ninguna aplicación específica. En tal caso, la utilidad $u\{p(\theta); \theta_0\}$, asociada a la distribución $p(\theta)$ cuando θ_0 es el verdadero valor de θ , depen-

derá de $p(\theta)$ únicamente a través de la (densidad de) probabilidad $p(\theta_0)$ asociada por tal distribución al verdadero valor θ_0 de la magnitud de interés. Se dice en este caso que la función de utilidad es *local*. Con una función de utilidad local, las distribuciones $p_1(\theta)$ y $p_2(\theta)$ de la figura comentada al principio de esta sección tendrían la misma utilidad si el verdadero valor de θ fuese θ_1 .

Mencionaremos sin demostración el siguiente resultado (Bernardo, 1979 a).

TEOREMA 7.2.1. Toda función de utilidad propia, local e invariante frente a transformaciones monótonas de θ debe ser de la forma

$$u\{p(\theta), \theta_0\} = A \log\{p(\theta_0)/p_0(\theta_0)\} \quad (1)$$

donde A es una constante cualquiera y $p_0(\theta)$ una distribución arbitraria tomada como origen.

Como consecuencia de este resultado, puede argumentarse que (2) es la única descripción satisfactoria de las preferencias que un científico debería tener entre las conclusiones de un problema de investigación puro. Al estudiar el diseño de experimentos, en la Sección 7.6, encontraremos nuevos argumentos en que apoyar esta afirmación.

7.3. Evaluación de utilidades

Según vimos en el capítulo de Fundamentos, los axiomas de coherencia establecen la necesidad de cuantificar las preferencias del decisor entre las posibles consecuencias de sus acciones mediante una función de utilidad u , de forma que $u(d, \theta)$ mida la deseabilidad de la consecuencia que se derivaría si sucediese θ y se hubiese tomado la decisión d : cuanto mayor sea este número más atractiva sería para el decisor la consecuencia a que se reflere.

Si las consecuencias que pueden derivarse de las distintas decisiones son naturalmente expresables de forma *cuantitativa* en una unidad común, entonces la función de utilidad será una función de tal magnitud y el problema se simplifica considerablemente. Es posible, en efecto, que todas las consecuencias relevantes a un problema de decisión médico puedan ser expresadas en términos de los años de vida que pueda esperar vivir el paciente; en un problema económico, es probable que todas las consecuencias puedan ser expresadas en términos monetarios. La función de utilidad será en estos casos una función *monótona* (pero no necesariamente *lineal*) de la correspondiente unidad común; maximizar la utilidad esperada equivaldrá a maximizar el valor de una función creciente de la esperanza de vida en el primer caso, o de una función creciente del beneficio esperado en el segundo.

No siempre existe, sin embargo, la suficiente homogeneidad entre las consecuencias como para poder establecer de forma natural una unidad común. Caben entonces dos alternativas: la primera es proceder a una comparación directa; la segunda es buscar su equivalente en una unidad común, frecuentemente dinero. Estudiaremos consecutivamente ambas posibilidades.

La comparación directa se hace en los términos descritos al definir la función de utilidad a partir de los axiomas de coherencia. Así pues, se escogen la mejor c^* y la peor c_* de las consecuencias posibles a las que se asigna arbitrariamente los valores uno y cero, respectivamente, de forma que

$$c_* \leq c \leq c^* \quad u(c_*) = 0, \quad u(c^*) = 1$$

Estas consecuencias de referencia c_* y c^* pueden pertenecer o no al conjunto de las consecuencias posibles; todo depende de cuales sean las consecuencias con las que sea más fácil establecer comparaciones, puesto que la sencillez de comparación debe ser el criterio que presida su elección.

Para cualquier otra consecuencia c se presentará al decisor la opción $\{c^*|p, c_*|(1-p)\}$ que le permite obtener c^* con probabilidad p o c_* con la probabilidad complementaria, $1-p$. Mediante prueba y error, modificando convenientemente el valor de p , se obtendrá un valor para el que la consecuencia y la opción son igualmente deseables, es decir un valor p_0 tal que

$$c \sim \{c^*|p_0, c_*|(1-p_0)\}$$

la utilidad de la consecuencia c es entonces el valor p_0 así determinado, esto es $u(c) = p_0$.

El método puede repetirse con otras consecuencias de referencia. Si el decisor es consistente las utilidades asignadas deberán ser una función lineal de las originales. En general, debido a las inconsistencias del decisor, la relación no será exactamente lineal; sin embargo, el valor medio de las utilidades así obtenidas, una vez referidas todas a una *escala común*, proporcionará generalmente una estimación suficientemente precisa de las utilidades buscadas.

Como ya hemos mencionado, un método alternativo consiste en referir todas las consecuencias a una utilidad común, frecuentemente dinero. Suele objetarse que esto no siempre puede hacerse, en el sentido de que parece difícil asignar un valor monetario a las consecuencias de una enfermedad o al placer de asistir a un concierto. Sin embargo, el hecho de que se suscriban seguros libres de enfermedad o se paguen determinados precios por las entradas de un concierto, sugiere que existe un cierto equivalente económico para la mayor parte de las consecuencias.

Si resultase posible determinar el equivalente monetario (positivo o negativo) de todas las consecuencias que intervienen en un problema de decisión, el problema de asignar utilidades quedaría reducido al de determinar una función que describa la utilidad del dinero.

Naturalmente, una primera aproximación consistiría en suponer que la utilidad del dinero es proporcional a su cantidad. Aunque esta hipótesis de *linealidad* puede proporcionar una aproximación válida cuando se manejan cantidades de dinero que resultan pequeñas comparadas con el capital total de que se dispone, se trata de una hipótesis claramente insostenible fuera de estos límites; obviamente, 100.000 pesetas pueden ser una cantidad importante para un obrero y una trivialidad para un terrateniente.

Del análisis anterior se desprende que la función de utilidad del dinero debe tener en cuenta los recursos que se poseen; denotaremos con $u(c)$ la utilidad de disponer de un *capital total* c . En consecuencia, la utilidad de obtener un *beneficio neto* b será $u(c+b) - u(c)$.

Una consideración detallada del significado de $u(c)$ nos permite sugerir que la función de utilidad del dinero debe cumplir las siguientes condiciones

- (i) $u(c) \geq 0$, $u(0) = 0$, $u(\infty) = 1$. (*positiva y acotada*)
- (ii) $u'(c) > 0$ (*creciente*)
- (iii) $u''(c) < 0$ (*cóncava*)

En efecto, parece razonable suponer que la utilidad de disponer de un capital total c es positiva y acotada; que crece con c , pero que lo hace de forma cada vez menos pronunciada; el incremento de utilidad producido por una determinada ganancia es positivo, pero tanto menor cuando más dinero se tiene.

Las condiciones (i), (ii) y (iii) exigidas a la función $u(c)$ que describe la utilidad de disponer de un capital total c , permiten analizar su forma. En efecto, en virtud de (iii), para todo c y para todo b existirá una probabilidad $p > 1/2$ tal que la opción $\{b|p, -b|(1-p)\}$, esto es la apuesta que permite ganar b con probabilidad p o perder b con la probabilidad complementaria, sea equivalente a la de quedarse en la situación inicial, con un capital c . En consecuencia, para ese valor de p ,

$$u(c+b)p + u(c-b)(1-p) = u(c)$$

A la diferencia $p-0,5$ se le llama la *prima probabilística* que el decisor exige para decidirse a aceptar una opción con esperanza monetaria nula, debido a la *aversión al riesgo* que implica la concavidad de la función de utilidad.

Es fácil comprobar que la prima probabilística puede escribirse como

$$p - \frac{1}{2} = \frac{2u(c) - u(c+b) - u(c-b)}{2u(c+b) - 2u(c-b)}$$

y, por lo tanto,

$$\lim_{c \rightarrow \infty} \frac{(p-0,5)}{2} \propto \frac{-u'(c)}{u(c)} = r(c) > 0$$

es una medida de la *aversión local al riesgo* que caracteriza al decisor. Parece razonable suponer que $r(c)$ es una función no decreciente de c , esto es que la aversión al riesgo permanece constante o disminuye al aumentar el capital.

Si $r(c) = a$ es constante, la solución de la correspondiente ecuación diferencial, esto es, de

$$au'(c) + u''(c) = 0, \quad a > 0$$

demuestra que la función de utilidad debe ser del tipo

$$u(c) = 1 - \exp(-\gamma c), \quad \gamma = 1/a$$

de forma que el logaritmo de la pérdida $1 - u(c)$ es lineal en c .

Existe, sin embargo, una amplia clase de funciones de utilidad con $r(c)$ decreciente. Pratt (1964) estudia algunas de ellas. Lindley (1971 b) sugiere utilizar una combinación lineal de dos utilidades con aversión al riesgo constante, lo que da lugar a una función de utilidad con aversión al riesgo decreciente.

La familia de funciones de utilidad definida por

$$u(c|\alpha, \beta, \gamma) = 1 - \alpha \exp(-\beta c) - (1 - \alpha) \exp(-\gamma c) \quad (1)$$

donde $c > 0$, $0 \leq \alpha \leq 1$, $\beta > 0$ y $\gamma > 0$, cumple las condiciones (i), (ii) y (iii) y describe una situación en la que la aversión al riesgo, siempre presente, disminuye a medida que aumenta el capital c de que se dispone. Se trata además de una clase de funciones analíticamente tratable, fácil de interpretar y que describe, al hacer variar sus parámetros α , β y γ , una amplia gama de posibilidades.

Si las preferencias del decisor pueden ser descritas mediante un elemento de la clase (1), el problema se reduce a determinar los valores de α , β y γ . Una forma de hacerlo es preguntar al decisor por el valor mínimo de p con el que aceptaría la opción $\{a|p, -b|(1-p)\}$, esto es una apuesta en la que ganaría a con probabilidad p o perdería b con la probabilidad complementaria. Con estos datos y conocido el capital total c de que dispone el decisor, podemos formular la ecuación

$$u(c+a)p + u(c-b)(1-p) = u(c)$$

Obviamente, para cada conjunto de valores (c, a, b, p) tendremos una ecuación que debe satisfacer u . Con tres de estas ecuaciones tendremos un sistema cuya solución dará los valores de α , β y γ . El procedimiento puede repe-

irse con distintos sistemas de ecuaciones. Debido a las inconsistencias del decisor, las soluciones no coincidirán exactamente pero su valor medio constituirá generalmente un buen estimador de los valores α , β y γ que describen las utilidades monetarias del decisor.

Ejemplo 7.3.1. Utilidades monetarias

Una persona que dispone de unos bienes totales valorados en 1.500.000 pesetas, contrata por 10.000 ptas. un seguro contra robo para un coche valorado en 500.000 ptas., pensando que, de otra manera, tendría probabilidad 0,014 de perder el coche. Acepta además participar en un negocio en el que puede ganar 25.000 ptas. con probabilidad 0,15 o perder 4.000 con la probabilidad complementaria y, finalmente, compra un paquete de acciones, que pueden hacerle ganar o perder 200.000 ptas. según se comporte el mercado de valores, al convencerse de que la probabilidad de ganar las 200.000 ptas. es 0,54. Determinar una función de utilidad del tipo (1) compatible con este comportamiento.

La función de utilidad $u(c)$ debe satisfacer (midiendo c en miles de pesetas) que

$$u(1.490) = u(1.000)0,014 + u(1.500)0,986$$

$$u(1.500) = u(1.525)0,15 + u(1.496)0,85$$

$$u(1.500) = u(1.700)0,54 + u(1.300)0,46$$

y puesto que $u(c)$ debe ser de la forma

$$u(c) = 1 - \alpha \exp(-\beta c) - (1 - \alpha) \exp(-\gamma c)$$

tenemos un sistema de tres ecuaciones con las tres incógnitas α , β y γ . Su solución, obtenida por prueba y error, es aproximadamente $\alpha = 0,5$, $\beta = 0,001$ y $\gamma = 0,0005$. En consecuencia, si se verifican las condiciones (i), (ii) y (iii) mencionadas en el texto, la utilidad que para esa persona tienen x miles de pesetas puede ser descrita mediante la función

$$u(x) = 1 - 0,5 \exp(-x/1.000) - 0,5 \exp(-x/2.000)$$

7.4. Valor esperado de la información

Consideremos un problema de decisión en el que el espacio de decisiones es D , el espacio de sucesos inciertos Θ y la función de utilidad $u(d, \theta)$. Sea $p(\theta)$ la distribución de probabilidad que, en un momento dado, describe la información de que se dispone sobre el valor de θ . La decisión óptima en ese momento es la que maximiza el valor de la utilidad esperada $u^*(d)$ definida por

$$u^*(d) = \sum u(d, \theta_i) p(\theta_i) \quad (\text{si } \theta \text{ es discreta}).$$

$$u^*(d) = \int u(d, \theta) p(\theta) d\theta \quad (\text{si } \theta \text{ es continua})$$

Consecuentemente, la utilidad que puede esperarse si se toma la decisión óptima es

$$u_0^* = \sup_{d \in D} u^*(d) \quad (1)$$

Supongamos ahora que, antes de tomar la decisión, se realiza un experimento ε y se obtiene como resultado unos datos z que contienen cierta información sobre el valor de θ . Después de obtener z , la información de que se dispone sobre el valor de θ vendrá descrita por la correspondiente distribución final $p(\theta|z)$; la decisión óptima será ahora la que maximice la correspondiente utilidad esperada $u^*(\varepsilon, z, d)$ definida por

$$u^*(\varepsilon, z, d) = \sum u(d, \theta_i) p(\theta_i|z) \quad (\text{si } \theta \text{ es discreta})$$

$$u^*(\varepsilon, z, d) = \int u(d, \theta) p(\theta|z) d\theta \quad (\text{si } \theta \text{ es continua})$$

Por lo tanto, la utilidad que puede esperarse si se toma la decisión óptima después de observar z será

$$u^*(\varepsilon, z) = \sup_{d \in D} u^*(d, \varepsilon, z), \quad (2)$$

de forma que el *valor esperado de la información* proporcionada por los datos z , esto es el incremento de utilidad que se puede esperar adoptando siempre la decisión óptima en cada caso, vendrá dada por

$$\nabla u^*(\varepsilon, z) = u^*(\varepsilon, z) - u_0^* \quad (3)$$

El valor esperado de la información proporcionada por el experimento ε antes de conocer su resultado será obviamente

$$\begin{aligned} \nabla u^*(\varepsilon) &= \int \nabla u^*(\varepsilon, z) p(z) dz = \\ &= \int u^*(\varepsilon, z) p(z) dz - u_0^* = u^*(\varepsilon) - u_0^* \end{aligned} \quad (4)$$

Si denotamos con ε_0 al *experimento vacío*, de forma que $u^*(\varepsilon_0)$ es la utilidad esperada de tomar la decisión basándose únicamente en la información inicial, tenemos obviamente que $u^*(\varepsilon_0) = u_0^*$. En el extremo opuesto, si denotamos

con ε_∞ a un experimento que nos proporcionase información *completa* sobre el valor de θ , tenemos que

$$u^*(\varepsilon_\infty) = \sum p(\theta_i) \sup_d u(d, \theta_i) \quad (\text{si } \theta \text{ es discreta})$$

$$u^*(\varepsilon_\infty) = \int p(\theta) \sup_d u(d, \theta) d\theta \quad (\text{si } \theta \text{ es continua})$$

En efecto, con información perfecta, se tomaría para cada θ_0 la decisión que maximice en D el valor de $u(d, \theta_0)$ de forma que, *antes* de disponer de esa información perfecta, la utilidad esperada de conseguirla viene dada por la expresión anterior.

El *valor esperado de la información perfecta* será consecuentemente

$$\nabla u^*(\varepsilon_\infty) = u^*(\varepsilon_\infty) - u^*(\varepsilon_0)$$

y puede demostrarse, como era de esperar, que para todo experimento ε

$$\nabla u^*(\varepsilon) \leq \nabla u^*(\varepsilon_\infty)$$

es decir, que el valor esperado de la información proporcionada por un experimento es siempre menor o igual que el valor esperado de la información perfecta. Este valor esperado de la información perfecta representa pues la cantidad máxima, en unidades de utilidad, que sería razonable pagar por obtener información adicional sobre el parámetro de interés.

Ejemplo 7.4.1. Valor diagnóstico de un test

El equipo médico que atiende a un determinado paciente considera que los síntomas que se observan pueden ser producidos por una y solo una de las enfermedades θ_1 , θ_2 v θ_3 , v se considera que $p(\theta_1) = p(\theta_2) = 2p(\theta_3)$. Se intenta determinar cual de ellas es la verdadera causa, recurriéndose para ello a determinar la cantidad X en mg de una determinada sustancia presente en la orina que el paciente elimina cada día. Investigaciones anteriores permiten suponer que

$$p(x|\theta_1) = N(x|12, 5)$$

$$p(x|\theta_2) = N(x|18, 8)$$

$$p(x|\theta_3) = N(x|16, 5)$$

Se realiza el experimento ϵ que consiste en recoger y analizar durante tres días consecutivos la orina del paciente, y se observan las cantidades 17, 18,5 y 17,5 mg de la sustancia estudiada. Suponiendo que el interés se centre únicamente en un diagnóstico correcto, de forma que la utilidad conseguida es uno o cero según el diagnóstico sea correcto o incorrecto, determinar el valor esperado de la información proporcionada por el test y el valor que podría esperarse de la información perfecta.

La distribución inicial de θ es obviamente $p(\theta) = (0,4, 0,4, 0,2)$. La utilidad esperada de la decisión θ_i es

$$u^*(\theta_i) = \sum_{j=1}^3 u(\theta_i, \theta_j) p(\theta_j) = p(\theta_i)$$

de forma que la utilidad esperada de cualquiera de las dos decisiones óptimas (θ_1 o θ_2) es $u^*(\theta_0) = 0,4$.

La verosimilitud de los datos $z = \{17, 18, 5, 17,5\}$ obtenidos como resultados del experimento ϵ viene dada por

$$p(z|\theta_1) = \Pi N(x_i|12, 5) = 7,23 \times 10^{-3}$$

$$p(z|\theta_2) = \Pi N(x_i|18, 8) = 1,23 \times 10^{-4}$$

$$p(z|\theta_3) = \Pi N(x_i|16, 5) = 4,20 \times 10^{-4}$$

v consecuentemente, utilizando el Teorema de Bayes, la distribución final de θ será $p(\theta|z) = (0,178, 0,303, 0,518)$. La utilidad esperada de la decisión θ_i es ahora

$$u^*(\epsilon, z, \theta_i) = \sum_{j=1}^3 u(\theta_i, \theta_j) p(\theta_j|z) = p(\theta_i|z)$$

de forma que la utilidad esperada de la decisión óptima (θ_3) es $u^*(\epsilon, z) = 0,518$. Consecuentemente, el valor de la información proporcionada por los datos z es $\nabla u^*(\epsilon, z) = 0,518 - 0,4 = 0,118$.

El valor esperado de la información perfecta sería

$$\begin{aligned} \nabla u^*(\epsilon) &= u^*(\epsilon_{\infty}) - u^*(\epsilon_0) = \\ &= \sum p(\theta_j) \sup_{\theta_i} u(\theta_i, \theta_j) - u^*(\epsilon_0) = \\ &= \sum p(\theta_j) - u^*(\epsilon_0) = 1 - u^*(\epsilon_0) = 1 - 0,4 = 0,6 \end{aligned}$$

Dada la forma especialmente sencilla que la función de utilidad adopta en este problema, el valor esperado de la información resulta ser una *probabilidad*. Así, la información espe-

perimental aumenta en 0,118 la probabilidad asociada a la enfermedad más probable (que pasa de ser 0,4 a ser 0,518), mientras que la información perfecta aumentaría en 0,6 esta probabilidad, que pasaría de 0,4 a la unidad (conocimiento perfecto).

En la Sección 7.2 vimos como un problema de inferencia estadística sobre el valor de θ podía ser formulado como un problema de decisión en el que el espacio de decisiones es el de las distribuciones de probabilidad de θ y la función de utilidad de la forma

$$u\{p(\theta), \theta_0\} = A \log \{p(\theta_0)/p(\theta)\} \quad (5)$$

Cuando en la definición de la función de utilidad se escoge la distribución inicial como distribución origen, resulta que el *valor* de la información proporcionada por los resultados de un experimento ϵ en un problema de inferencia es proporcional a la *cantidad* de información obtenida.

En efecto, de acuerdo con (1)

$$\begin{aligned} u^*(\epsilon_0) &= \frac{\sup}{p(\theta)} \int [A \log \{q(\theta)/p(\theta)\}] p(\theta) d\theta = \\ &= A \int q(\theta) \log \{q(\theta)/p(\theta)\} d\theta = 0 \end{aligned}$$

puesto que (5) es una función de utilidad propia, y por lo tanto su valor esperado es máximo cuando $q(\theta) = p(\theta)$.

Análogamente, utilizando (2)

$$\begin{aligned} u^*(\epsilon, z) &= \frac{\sup}{q(\theta)} \int [A \log \{q(\theta)/p(\theta)\}] p(\theta|z) d\theta = \\ &= A p(\theta|z) \log \{p(\theta|z)/p(\theta)\} d\theta \end{aligned}$$

En consecuencia, comparando con la Definición 5.4.1, la utilidad esperada de realizar un experimento ϵ , obtener unos datos z , y utilizar la distribución final $p(\theta|z)$ para describir el resultado es

$$u^*(\epsilon, z) = A I^0\{\epsilon, p(\theta|z)\}$$

y por tanto

$$\nabla u^*(\epsilon, z) = u^*(\epsilon, z) - u^*(\epsilon_0) = A I^0\{\epsilon, p(\theta|z)\}$$

como queríamos demostrar.

Por otra parte, *cualquiera* que sea la distribución inicial elegida como origen, el *valor* esperado de la información proporcionada por un experimen-

to ϵ en un problema de inferencia es proporcional a la cantidad de información que puede esperarse de él.

En efecto, utilizando (1) y (2), resulta

$$u^*(\epsilon_0) = A \int p(\theta) \log \{p(\theta)/p_0(\theta)\} d\theta$$

$$u^*(\epsilon, z) = A \int p(\theta|z) \log \{p(\theta|z)/p_0(\theta)\} d\theta$$

y, por tanto,

$$\nabla u^*(\epsilon) = \int \nabla u^*(\epsilon, z) p(z) dz =$$

$$= \int u^*(\epsilon, z) p(z) dz - u^*(\epsilon_0)$$

$$= A \int p(z) \int p(\theta|z) \log \{p(\theta|z)/p_0(\theta)\} d\theta$$

de forma que, de acuerdo con la Definición 5.4.2,

$$\nabla u^*(\epsilon) = A I^0\{\epsilon, p(\theta)\}$$

Como consecuencia de este resultado, en un problema de investigación pura, el mejor experimento es el que maximiza la información esperada sobre el parámetro de interés. El contenido intuitivo de este resultado proporciona una nueva justificación indirecta para la definición de información adoptada en el Capítulo 5.

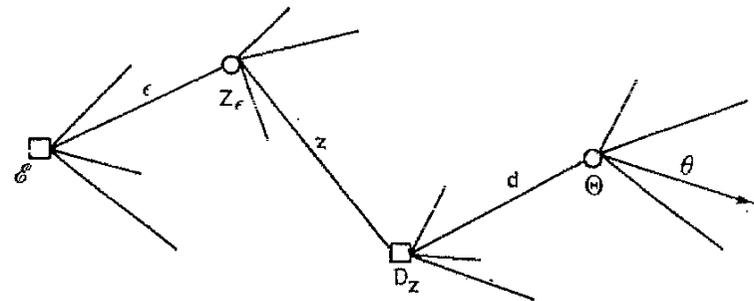
En el análisis realizado hasta ahora, no hemos considerado el coste del experimento; nos hemos limitado a calcular el valor esperado de la información que nos proporciona. Como hemos visto, este valor viene medido por el incremento en la utilidad esperada de tomar la decisión óptima, conseguido gracias a la información adicional obtenida sobre el parámetro de interés. Sin embargo, para elegir de forma razonable entre un conjunto de experimentos posibles es obviamente necesario incluir en el análisis el coste del experimento; este es el objeto de la próxima sección.

7.5. Diseño de experimentos

En los problemas de inferencia que hemos analizado hasta ahora hemos supuesto que el experimento había sido escogido ya y nos limitábamos a ana-

lizar la información contenida en sus resultados. Abordaremos ahora el problema de elegir en un experimento concreto de la familia \mathcal{E} de experimentos posibles.

De forma totalmente general, puesto que hemos demostrado que los problemas de inferencia son casos particulares de problemas de decisión, podemos plantear el problema de la elección del experimento como un problema de decisión con dos decisiones sucesivas. En efecto, en primer lugar debe elegirse un experimento determinado $\epsilon \in \mathcal{E}$ y a continuación, a la vista de los resultados obtenidos z , debe elegirse la decisión $d \in D_z$ que maximiza la utilidad esperada entre las decisiones D_z que resultan posibles a la vista de los resultados z ; se obtiene entonces una consecuencia c que depende del suceso incierto θ que haya tenido lugar (ver figura)



Naturalmente, la utilidad de este resultado final será de la forma $u(c) = u(\epsilon, z, d, \theta)$.

Frecuentemente, $u(\epsilon, z, d, \theta)$ admite una descomposición aditiva de forma que la utilidad del resultado final c puede calcularse como la diferencia entre la utilidad terminal $u(d, \theta)$ de tomar la decisión d cuando sucede θ y el coste $c(\epsilon, z)$ de realizar el experimento ϵ cuando se obtiene el resultado z . En este caso, podemos escribir

$$u(\epsilon, z, d, \theta) = u(d, \theta) - c(\epsilon, z) \tag{1}$$

Naturalmente, las cantidades $u(d, \theta)$ y $c(\epsilon, z)$ deben estar medidas en las mismas unidades.

Procediendo de derecha a izquierda, como en todos los árboles de decisión (ver Sección 3.5), la utilidad esperada de cada una de las ramas del nodo de decisión D_z (ver figura) es

$$u^*(\varepsilon, z, d) = \sum u(\varepsilon, z, d, \theta_i) p(\theta_i|z) \quad (\text{si } \theta \text{ es discreta})$$

$$u^*(\varepsilon, z, d) = \int u(\varepsilon, z, d, \theta) p(\theta|z) d\theta \quad (\text{si } \theta \text{ es continua})$$

De estas ramas, debe elegirse la de mayor utilidad esperada; en consecuencia, la utilidad esperada asociada al nodo de decisión D_2 será

$$u^*(\varepsilon, z) = \sup_{d \in D_2} u^*(\varepsilon, z, d)$$

Análogamente, la utilidad esperada de cada una de las ramas del nodo de decisión E será

$$u^*(\varepsilon) = \sum u^*(\varepsilon, z_i) p(z_i) \quad (\text{si } z \text{ es discreta})$$

$$u^*(\varepsilon) = \int u^*(\varepsilon, z) p(z) dz \quad (\text{si } z \text{ es continua})$$

De estos valores debe elegirse el de mayor utilidad esperada que corresponde, por definición, al *experimento óptimo* ε^+ ; de forma que

$$u^* = \sup_{\varepsilon \in \mathcal{E}} u^*(\varepsilon) = u^*(\varepsilon^+)$$

En la clase \mathcal{E} de experimentos posibles debe ser siempre incluido el *experimento vacío* ε_0 para incluir la posibilidad de que resulte óptimo tomar la decisión basándose únicamente en la información inicial.

El *valor esperado del experimento* ε , es decir el incremento de utilidad que puede esperarse por realizarlo, viene dado por

$$v^*(\varepsilon) = u^*(\varepsilon) - u^*(\varepsilon_0)$$

Naturalmente, puesto que $u^*(\varepsilon_0)$ es una constante, el experimento con mayor valor esperado es el experimento óptimo ε^+ . Si la función de utilidad puede descomponerse en la forma (1), resulta que el valor esperado del experimento es igual al valor esperado de la información que proporciona menos su coste esperado.

Una vez encontrado el experimento óptimo ε^+ , esto es, una vez resuelto el problema del *diseño del experimento*, se procede a realizarlo y a determinar la decisión óptima d^+ dentro del conjunto D_2 de las decisiones que resultan posibles una vez observado el resultado z del experimento elegido ε^+ . Esta decisión óptima d^+ es naturalmente aquella que maximiza la correspondiente utilidad esperada.

$$u^*(\varepsilon^+, z, d^+) = \sup_{d \in D_2} u^*(\varepsilon^+, z, d)$$

Un ejemplo de este proceso fue ya adelantado al estudiar, en general, el análisis secuencial de decisiones (Ejemplo 3.5.2).

Ejemplo 7.5.1. Exploraciones peligrosas

A un paciente que presenta síntomas compatibles con las enfermedades θ_1 , θ_2 y θ_3 le pueden ser aplicados los tratamientos t_1 o t_2 de acuerdo con la siguiente tabla de utilidades, expresada en años esperados de vida

	θ_1	θ_2	θ_3
t_1	20	2	5
t_2	4	14	6

la información inicial que se dispone sobre la verdadera enfermedad del paciente puede describirse mediante la distribución $p(\theta) = \{0,3, 0,3, 0,4\}$. Esta información puede mejorarse mediante una exploración interna en la que el paciente tiene una probabilidad 0,9 de sobrevivir. La información obtenida en caso de realizar la exploración puede ser de dos tipos, A y B y se sabe que

$$p(A|\theta_1) = 0,95$$

$$p(A|\theta_2) = 0,02$$

$$p(A|\theta_3) = 0,40$$

Determinar la estrategia óptima.

Se trata de elegir entre dos posibles «experimentos», ε_0 : decidir sin exploración alguna, y ε_1 : realizar la exploración y decidir entonces.

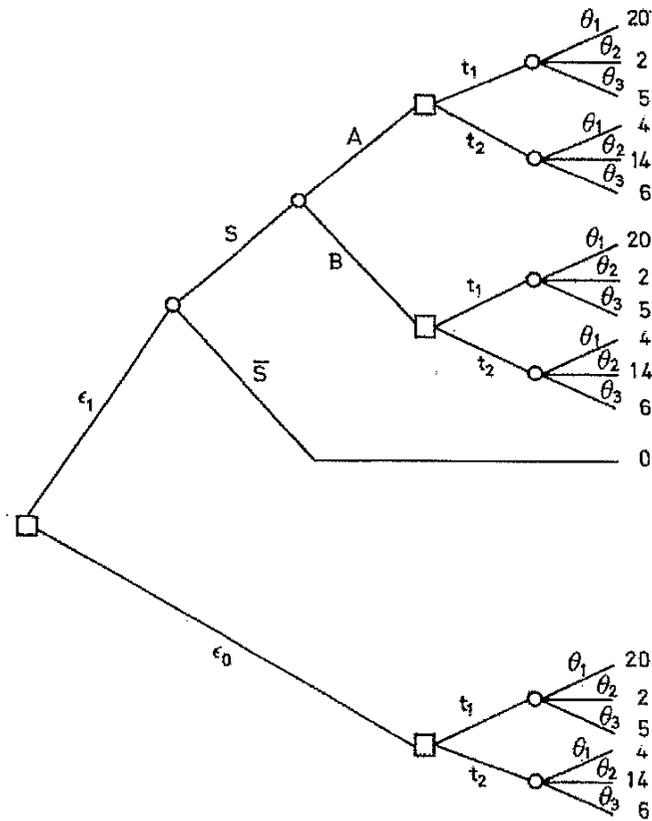
Si no se realiza la exploración, la utilidad esperada de cada uno de los dos tratamientos es

$$u^*(\varepsilon_0, t_1) = \sum u(t_1, \theta_i) p(\theta_i) = 8,6$$

$$u^*(\varepsilon_0, t_2) = \sum u(t_2, \theta_i) p(\theta_i) = 7,8$$

de forma que la utilidad esperada de aplicar el tratamiento óptimo sin información adicional, que resulta ser t_1 , es $u^*(\varepsilon_0) = 8,6$.

Si se realiza la exploración y el resultado es A , la distribución final de θ es, haciendo uso del Teorema de Bayes, $p(\theta|A) = (0,632, 0,013, 0,355)$.



y la utilidad esperada de cada uno de los tratamientos, es entonces

$$u^*\{\epsilon_1, S, A, t_1\} = \sum u\{\epsilon_1, \theta_i\} p(\theta_i|A) = 14,4$$

$$u^*\{\epsilon_1, S, A, t_2\} = \sum u\{\epsilon_1, \theta_i\} p(\theta_i|A) = 4,8$$

donde S indica el suceso de que el paciente ha sobrevivido a la exploración. En consecuencia, la utilidad esperada de aplicar el tratamiento óptimo, t_1 en este caso, cuando el resultado de la exploración es A resulta ser $u^*\{\epsilon_1, S, A\} = 14,4$. Análogamente, la distribución final de θ si el resultado de la exploración es B es $p(\theta|B) = (0,27, 0,536, 0,437)$ y la utilidad esperada de aplicar el tratamiento óptimo, que ahora sería t_2 , si el resultado de la exploración es B, resulta ser $u^*\{\epsilon_1, S, B\} = 10,2$.

La distribución predictiva de los resultados de la exploración resulta ser $p(A|S) = 0,451$,

$p(B|S) = 0,549$ y en consecuencia, la utilidad esperada de realizar la exploración, si el paciente sobrevive es

$$u^*\{\epsilon_1, S\} = u^*\{\epsilon_1, S, A\}p(A|S) + u^*\{\epsilon_1, S, B\} p(B|S) = 12,1$$

Como el paciente solo tiene probabilidad $p(S) = 0,9$ de sobrevivir la exploración, la utilidad esperada de decidir realizarla es

$$u^*\{\epsilon_1\} = u^*\{\epsilon_1, S\}p(S) + 0\{1 - p(S)\} = 10,9$$

que es mayor que la utilidad esperada $u^*\{\epsilon_0\}$ de no realizar la exploración.

Consecuentemente, la estrategia óptima es realizar la exploración; si el paciente sobrevive y el resultado de la exploración es A aplicar el tratamiento t_1 , y si el paciente sobrevive y el resultado es B aplicar t_2 .

Supongamos que en un experimento científico se valora en g la utilidad de una unidad de información sobre el parámetro de interés θ y que el coste de realizar n observaciones independientes de una cantidad aleatoria X cuya densidad de probabilidad es $p(x|\theta)$ es de la forma $c(n) = c_0 + nc$. La utilidad esperada de realizar n observaciones será entonces de la forma

$$u^*(n) = gI^g\{n, p(\theta)\} - c_0 - nc$$

donde $p(\theta)$ es la distribución inicial de θ . El cálculo exacto de $I^g\{n, p(\theta)\}$ es complicado pero; utilizando los resultados sobre distribuciones asintóticas, puede demostrarse (Bernardo, 1979 c).

TEOREMA 7.5.1. Si $\psi = \psi(\theta)$ es una transformación monótona de θ con densidad aproximadamente normal de varianza σ_0^2 entonces para valores de n no muy pequeños

$$I^g\{n, p(\theta)\} \approx \frac{1}{2} \log \left(1 + \frac{\sigma_0^2}{\sigma_x^2} \right)$$

donde

$$\frac{1}{\sigma_x^2} = \int \frac{b(\theta)}{(\partial\psi/\partial\theta)^2} p(\theta)d\theta$$

$$b(\theta) = \int p(x|\theta) \frac{\partial^2}{\partial\theta^2} \log p(x|\theta) dx$$

En particular, si $p(x|\theta) = N(x|\mu, \sigma)$ y $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$, resulta $\sigma_x = \sigma$ y

$$I^g\{n, N(\mu|\mu_0, \sigma_0^2)\} = \frac{1}{2} \log \left(1 + n \frac{\sigma_0^2}{\sigma^2} \right)$$

tratándose, en este caso, de un resultado exacto.

Utilizando el teorema anterior, la utilidad esperada de realizar n observaciones será, aproximadamente,

$$u^*(n) \approx \frac{g}{2} \log \left(1 + n \frac{\sigma_0^2}{\sigma_x^2} \right) - c_0 - nc$$

expresión que alcanza su máximo cuando $n = n'$ donde

$$n' = \frac{1}{2} \frac{g}{c} - \frac{\sigma_0^2}{\sigma_x^2} \quad (2)$$

El tamaño muestral óptimo será pues, aproximadamente, el entero positivo n^* más próximo a n' , si $u^*(n^*) > 0$ y cero si $u^*(n^*) \leq 0$.

Ejemplo 7.5.2. Tamaño óptimo de una encuesta

La información inicial sobre la proporción θ de elementos de la población con una determinada característica puede ser descrita por la distribución $Be(\theta|5, 3)$ y se estaría dispuesto a pagar 10.000 ptas. por saber si θ es o no es mayor que 0,5. El coste de un sondeo es de 5.000 ptas. fijas, más 100 pesetas por encuesta realizada. Determinar aproximadamente el tamaño óptimo de la muestra.

La probabilidad p asociada al suceso $\theta < 0,5$ será, utilizando la correspondiente transformación normalizadora (Teorema 6.3.2)

$$p = P[\theta < 0,5] = \Phi \left[\log \frac{\theta}{1-\theta} < 0 \right] = \Phi(-\mu_0/\sigma_0)$$

donde, si $p(\theta) = Be(\theta|\alpha, \beta)$,

$$\mu_0 = \log \frac{\alpha}{\beta} + \frac{\alpha - \beta}{2\alpha\beta}$$

$$\sigma_0 = \sqrt{\{(\alpha + \beta)/\alpha\beta\}}$$

En este caso, $\alpha = 5$, $\beta = 3$ y, por tanto, $\mu_0 = 0,577$, $\sigma_0 = 0,730$ y $p = 0,215$. En consecuencia, la información sobre θ que proporcionaría saber si $\theta < 0,5$ o $\theta > 0,5$ vendrá dado por (Teorema 5.4.2)

$$i - p \log p - (1-p) \log (1-p) = 0,52$$

Puesto que por esta información se pagarían 10.000 pesetas, el valor de la unidad de información resulta ser $g = 10.000/0,52 = 19.236$.

Para calcular la información proporcionada sobre θ por un sondeo de tamaño n utilizaremos el Teorema 7.5.1 con la transformación $\psi(\theta) = \log \{\theta/(1-\theta)\}$. Puesto que se trata de datos binomiales,

$$p(x|\theta) = \theta^x(1-\theta)^{n-x}$$

$$h(\theta) = - \int p(x|\theta) \frac{\partial^2}{\partial \theta^2} \log p(x|\theta) dx = \frac{1}{\theta(1-\theta)}$$

$$\frac{\partial \psi}{\partial \theta} = \frac{1}{\theta(1-\theta)}$$

y, por tanto,

$$\frac{1}{\sigma_x^2} = \int \frac{h(\theta)}{(\partial \psi / \partial \theta)^2} p(\theta) d\theta = \int \theta(1-\theta) Be(\theta|\alpha, \beta) d\theta =$$

$$= E(\theta|\alpha, \beta) - E(\theta^2|\alpha, \beta) =$$

$$= \frac{\alpha}{\alpha + \beta} - \frac{(\alpha + 1)\alpha}{(\alpha + \beta)(\alpha + \beta + 1)} = \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \quad (3)$$

Consecuentemente, la información esperada sobre θ será

$$I^n(n, Be(\theta|\alpha, \beta)) = \frac{i}{2} \log \left\{ 1 + n \frac{\sigma_0^2}{\sigma_x^2} \right\} =$$

$$= \frac{1}{2} \log \left\{ 1 + \frac{n}{\alpha + \beta + 1} \right\}$$

puesto que, en virtud del Teorema 6.3.2,

$$\sigma_0^2 = D^2(\psi|\alpha, \beta) = \frac{(\alpha + \beta)}{\alpha\beta} \quad (4)$$

Introduciendo (3) y (4) en (2), el tamaño muestral óptimo es, pues, el entero más próximo a

$$n' = \frac{i}{2} \frac{g}{c} - (\alpha + \beta + 1) = 87,2,$$

si su utilidad esperada es positiva. Como

$$u^*(n) = \frac{g}{2} \log \left\{ 1 + \frac{n}{\alpha + \beta + 1} \right\} - c_0 - nc$$

resulta $u^*(87) = 9,065$ de forma que el tamaño óptimo de la muestra es, aproximadamente, de 87 unidades.

7.6. Algunos problemas médicos de decisión

La mayor parte de los problemas de decisión con los que un médico debe enfrentarse exigen un estudio específico detallado para su solución satisfactoria. Los conceptos que resultan necesarios para tal estudio han sido desarrollados a lo largo del texto de forma que el lector que los haya adquirido estará en condiciones de analizar adecuadamente cualquier problema específico de decisión, necesitando tan sólo la ayuda técnica de un matemático si el modelo utilizado da lugar a cálculos complicados.

Puede resultar útil, sin embargo, señalar las características fundamentales de los tres tipos de decisión médica más frecuente: calibrado, diagnóstico y elección de tratamiento.

7.6.1. Calibrado

Supongamos que se dispone de dos tratamientos alternativos t_1 y t_2 que inciden sobre una determinada magnitud X . El problema clásico de *calibrado* consiste en obtener una serie de medidas $\{x_{11}, x_{12}, \dots, x_{1n}\}$ de la magnitud X en elementos tratados con t_1 , obtener otra serie de medidas $\{x_{21}, x_{22}, \dots, x_{2m}\}$ en elementos tratados con t_2 y, a la vista de estos resultados, decidir si t_1 y t_2 tienen efectos realmente diferentes; en caso afirmativo, interesa medir la importancia de tal diferencia.

Si X es una magnitud continua, suele suponerse que las medidas $z_1 = \{x_{11}, \dots, x_{1n}\}$ y $z_2 = \{x_{21}, \dots, x_{2m}\}$ son muestras aleatorias de poblaciones normales (*) $N(x_1|\mu_1, \sigma_1)$ y $N(x_2|\mu_2, \sigma_2)$ respectivamente y se procede entonces a estudiar $\theta = \mu_1 - \mu_2$ esto es la diferencia de sus medias, o bien su cociente $\psi = \mu_1/\mu_2$. El problema de calibrado se reduce entonces a la obtención de la distribución final del parámetro de interés, esto es de $p(\theta|z_1, z_2)$ o de $p(\psi|z_1, z_2)$.

Si la distribución final θ tuviese su centro próximo a cero o la de ψ tuviese su centro próximo a uno, no existiría diferencia apreciable entre ambos tratamientos. En caso contrario $E(\theta|z_1, z_2)$ señalaría el orden de magnitud de la diferencia, en las unidades utilizadas para medir X , entre los resultados de aplicar ambos tratamientos. Análogamente, $E(\psi|z_1, z_2)$ indicaría el factor por el que el tratamiento t_1 multiplica el efecto del tratamiento t_2 .

Las fórmulas concretas de $p(\theta|z_1, z_2)$ y $p(\psi|z_1, z_2)$ dependen de la información inicial de que se disponga. Las distribuciones finales de *referencia* para θ y ψ que describen la información sobre sus valores contenida en los

(*) Si la hipótesis de normalidad no resultase aceptable podría procederse previamente a realizar una transformación normalizadora adecuada al problema.

datos experimentales, han sido obtenidas, respectivamente, por Sanjuan (1979) y Sendra (1979).

Para la diferencia de medias θ resulta, aproximadamente (*)

$$\pi(\theta|z_1, z_2) \approx S_f(\theta|\bar{x}_1 - \bar{x}_2, [\{s_1^2/(n-1) + s_2^2/(m-1)\}/b]^{1/2}, f) \quad (1)$$

donde

$$f = 4 + \frac{(a+b)^2}{a^2/(n-5) + b^2/(m-5)}$$

$$b^2 = \frac{f}{(f-2)} \frac{1}{(a+b)}$$

$$a = \frac{s_1^2/(n-1)}{s_1^2/(n-1) + s_2^2/(m-1)} \frac{n-1}{n-3}$$

$$b = \frac{s_2^2/(m-1)}{s_1^2/(n-1) + s_2^2/(m-1)} \frac{m-1}{m-3}$$

y donde

$$\bar{x}_1 = \Sigma x_{1i}/n, \quad \bar{x}_2 = \Sigma x_{2i}/m,$$

$$s_1^2 = \Sigma(x_{1i} - \bar{x}_1)^2/n \quad \text{y} \quad s_2^2 = \Sigma(x_{2i} - \bar{x}_2)^2/m$$

Si n y m son grandes, f es también grande, b se aproxima a la unidad y, consecuentemente, (1) puede aproximarse por

$$\pi(\theta|z_1, z_2) \approx N(\theta|\bar{x}_1 - \bar{x}_2, \sqrt{\{s_1^2/(n-1) + s_2^2/(m-1)\}})$$

Ejemplo 7.6.1. Análisis de sangre

Se sospecha que la sangre de los pacientes que padecen lesiones renales tiene un contenido potásico por debajo de los valores normales. Para comprobarlo, se analiza la sangre de 15 individuos normales obteniéndose en mg/100 ml, un valor medio de 168 y una desviación típica de 21, y la de 25 pacientes que padecen lesiones renales que dan lugar a un valor medio

(*) La solución exacta, más complicada, la recoge Sanjuan (1979). En la literatura clásica, este problema se conoce con el nombre de problema de *Behrens-Fisher*.

de 152, con una desviación típica de 35. Suponiendo que se trata de datos normales, determinar la probabilidad de que, efectivamente, el valor medio del contenido en potasio de la sangre de los pacientes con lesiones renales esté por debajo del correspondiente a individuos normales.

De acuerdo con (1), si μ_1 es la media de la distribución que describe el contenido potásico en la sangre de individuos normales, μ_2 la correspondiente a los individuos con lesiones renales, y $\theta = \mu_1 - \mu_2$, resulta $a = 0,446$, $b = 0,674$, $t = 33,45$, $b^2 = 0,950$ y

$$\pi(\theta|\bar{x}, \bar{s}, s) = N(\theta|16, 9,203, 33,45) = N(\theta|16, 9,203)$$

de forma que

$$p[\theta > 0] = 0,95$$

Los resultados correspondientes al cociente de medias, $\psi = \mu_1/\mu_2$, son mucho más complicados. Bernardo (1977) obtiene la distribución final de referencia de ψ en el caso particular en que $\sigma_1 = \sigma_2$; Sendra (1979) obtiene la distribución final de referencia en el caso general y desarrolla técnicas de integración numérica apropiadas para su estudio.

7.6.2. Diagnóstico

Un método de trabajo comúnmente utilizado en la práctica médica consiste en *diagnosticar* una determinada enfermedad θ entre el conjunto de enfermedades Θ que son compatibles con el vector $x \in X$ de *indicadores* que caracterizan al paciente, basándose en la comparación de x con el *banco de datos* $z = \{(x_i, \theta_i), i = 1, \dots, n\}$ disponible, formado por los indicadores x_i de n pacientes cuyas enfermedades θ_i han sido finalmente establecidas. Una vez diagnosticada la enfermedad el médico recomienda el tratamiento más adecuado para curarla. El vector $x = (r, s)$ de indicadores incluye tanto las *características propias* r del paciente (edad, sexo, ...) como el conjunto de *síntomas* s producidos por la enfermedad (hipertensión sanguínea, modificaciones en la composición de la orina, ...).

Las técnicas de *análisis discriminante* han sido utilizadas con este objeto. Mediante este procedimiento se determina cuál es la enfermedad θ más verosímil una vez observados los indicadores x y se actúa entonces asignando el tratamiento óptimo correspondiente a esa enfermedad *como si no existiese duda alguna* sobre la exactitud del diagnóstico. En general, se trata de un criterio de decisión *incorrecto*. En efecto, puede existir otro tratamiento t' que, aunque de eficacia algo menor que t si θ es la verdadera enfermedad que padece el paciente, sea mucho mejor que t si la verdadera enfermedad no es θ .

Es comúnmente aceptado que rara vez se puede *garantizar* la certeza de

un diagnóstico. Aparece así la necesidad de que la solución a un problema de diagnóstico adopte la forma de una distribución de probabilidad $p(\theta|x, z)$ que describa las probabilidades asignadas a las distintas enfermedades que puede padecer un paciente con vector de indicadores x , dada la información proporcionada por el banco de datos z . A esta distribución final de θ se la conoce con el nombre de *distribución diagnóstica*.

La construcción de la distribución diagnóstica es conceptualmente sencilla, basándose en el uso del Teorema de Bayes y del Teorema de la probabilidad total. Resulta, en efecto, que bajo ciertas hipótesis razonables (*)

$$p(\theta|x, z) \propto p(s|\theta, z) p(\theta|r)$$

$$p(s|\theta, z) = \int p(s|\theta, \omega) p(\omega|z, \theta) d\omega$$

$$p(\omega|z, \theta) \propto \prod_{i=1}^{n_\theta} p(s_i|\theta, \omega) p(\omega|\theta)$$

donde $p(\theta|r)$ es la probabilidad *inicial* de que el paciente padezca la enfermedad θ dadas sus características propias r , $p(s|\theta, \omega)$ es un modelo paramétrico que describe el comportamiento de los síntomas s en función de la enfermedad padecida θ y de un parámetro marginal desconocido ω , $p(\omega|\theta)$ es la distribución inicial condicional de ese parámetro marginal y n_θ el número de pacientes en el banco de datos que tienen la enfermedad θ .

Suele suponerse, tras las necesarias transformaciones, que $p(s|\theta, \omega)$ es un modelo normal multivariante. Las fórmulas (Bernardo, 1978) parecen complicadas pero el esfuerzo resulta recompensado. En efecto, Dombal *et al.* (1972) consiguen con métodos semejantes un 92 % de diagnósticos correctos sobre la causa de dolores abdominales agudos, mientras Knill-Jones *et al.* (1973) obtienen un 98 % de diagnósticos correctos sobre la causa de la ictericia.

7.6.3. Elección de tratamiento

Frecuentemente, el cálculo de la distribución diagnóstica $p(\theta|x, D)$ estudiada en el apartado anterior no es más que un paso intermedio para la elección del tratamiento óptimo.

La elección de tratamiento es en efecto un problema de decisión en el que el espacio de decisiones es el conjunto T de tratamientos posibles y

(*) Ver, por ejemplo, Aitchison & Dunsmore (1970, cap. 11), Bernardo (1978).

dónde, frecuentemente, el conjunto Θ de los sucesos inciertos relevantes es el de las posibles enfermedades que se considera posible que sufra el paciente.

Si $u(t, \theta)$ es la función de utilidad que mide en las unidades más adecuadas al problema (años de vida esperados, tiempo de permanencia en el hospital, coste, ...) las consecuencias de aplicar el tratamiento t cuando la enfermedad que padece el paciente es θ , entonces el tratamiento óptimo t^+ es obviamente el que maximiza la utilidad esperada

$$u^*(t) = \int u(t, \theta) p(\theta|x, z) d\theta$$

dónde $p(\theta|x, z)$ es la distribución diagnóstica; en efecto, esta distribución es precisamente la que recoge toda la información de que se dispone sobre θ en el momento de tomar la decisión.

En aquellos problemas en los que no existe una clasificación clara en enfermedades posibles, como pasa a menudo cuando se trata de trastornos psiquiátricos, resulta preferible describir los resultados de los distintos tratamientos mediante una función de utilidad de la forma $u(x, y)$ que describa la utilidad, en unidades adecuadas, de pasar del vector de indicadores x al vector de indicadores y como consecuencia del tratamiento. Naturalmente, el vector de indicadores final es una cantidad aleatoria cuya distribución dependerá del vector inicial x y del tratamiento elegido. En consecuencia, el tratamiento óptimo será el que maximice la utilidad esperada

$$u^*(t) = \int u(x, y) p(y|x, t) dy$$

La *distribución prognóstica* $p(y|x, t)$ que aparece en esta expresión puede calcularse mediante métodos parecidos a los descritos para la distribución diagnóstica (Bermudez, 1981).

7.7. Discusión y referencias

El tratamiento que hemos dado a los problemas de estimación puntual y de contraste de hipótesis difiere notablemente de su tratamiento como problema de inferencia en la metodología clásica; en efecto, desde el punto de vista adoptado en este libro, sólo tiene sentido plantearse estos problemas como problemas específicos de decisión: si se desea hacer inferencias sobre θ debe describirse *toda* la información de que se dispone sobre su valor y esta información *únicamente* la recoge la correspondiente distribución final.

Se ha argumentado a veces que la metodología que se deduce de la teoría de la decisión podría ser inapropiada para problemas de inferencia en los que, aparentemente, no hay que tomar decisión alguna. Hemos demostrado sin

embargo que la inferencia estadística puede ser descrita como un problema específico de decisión al que consecuentemente hay que aplicar la metodología que se deduce de los principios de coherencia. La idea de utilizar la familia de distribuciones finales del parámetro como espacio de decisiones resulta notablemente fértil, puesto que no sólo permite situar la inferencia estadística en el contexto de la teoría de la decisión sino que, como hemos visto, proporciona además una interpretación natural de las cantidades de información en términos de utilidades correspondientes a problemas de inferencia.

La descripción analítica de la función de utilidad del dinero facilita la especificación de utilidades en numerosos problemas concretos. Su concavidad, permite por otra parte demostrar que al maximizar la utilidad esperada en un problema económico se tiene en cuenta implícitamente el riesgo de la inversión. Así, por ejemplo, resulta generalmente óptimo *diversificar* una cartera de valores y no canalizar toda la inversión en aquellas acciones de las que se espera mayor rentabilidad, como sucedería si se utilizase una función lineal para describir las utilidades monetarias.

El problema de diseño de un experimento precede cronológicamente al análisis de sus resultados. Resulta no obstante necesario estudiar este tema en último lugar debido a que no puede decidirse cuál es el mejor experimento hasta haber decidido lo que se haría con sus eventuales resultados; en efecto, la elección del experimento más adecuado no es más que un caso particular de un problema de decisiones sucesivas y en estos problemas hay siempre que proceder por *inducción inversa* desde los resultados finales hacia las decisiones iniciales.

La Teoría de la Decisión es una disciplina relativamente moderna. El texto de Raiffa & Schlaifer (1961) constituye la primera exposición sistemática de los resultados a que conduce el principio de maximización de la utilidad esperada; todavía hoy sigue siendo un libro de consulta necesario debido a la enorme cantidad de material que contiene. La referencia moderna por excelencia, a nivel de monografía, es el libro de DeGroot (1970).

En los años setenta aparecieron diversos textos que, a nivel más elemental, desarrollaban las ideas básicas de teoría de la decisión. Citaremos entre ellos los de Aitchison (1980), LaValle (1970), Lindley (1971 b) y Winkler (1972).

Uno de los aspectos más importantes de la moderna teoría de la decisión es que sus conceptos fundamentales pueden ser comprendidos sin una preparación matemática elevada, lo que ha facilitado enormemente el desarrollo de sus aplicaciones en los campos más diversos. Aplicaciones específicas en Medicina han sido descritas, entre otros, por Aitchison & Dunsmore (1970), Betaque & Gorry (1971), Ginsberg (1970), Ginsberg & Offensend (1968), Gustafson *et al.* (1969), Lusted (1968) y Savage (1970).

PROBLEMAS

1. En un estudio dietético se determinan en g de proteínas por 100 g de materia comestible, el contenido proteínico de 50 partidas de ruidías blancas, encontrándose un valor medio muestral de 22 con una desviación típica de 5. Suponiendo que se trata de observaciones normales, contrastar la hipótesis de que el valor medio supera los 23 g, suponiendo que la función de utilidad es de la forma $u(H_1, H_2) = 1$ si $z = 1$ y cero en caso contrario.
2. Experimentos anteriores inclinan a un equipo de científicos a estimar en el 98 % la proporción de tumores que resultan destruidos por un determinado tipo de radiación y les permiten también asignar probabilidad 0,95 a la hipótesis de que dicho porcentaje es mayor del 95 %. Realizadas unas pruebas, resultan destruidos 162 de los 170 tumores tratados. Si la hipótesis es cierta, la utilidad de aceptarla es 1, y la de rechazarla 0,3; si es falsa, la utilidad de aceptarla es 0,1 y la de rechazarla 0,6. Determinar la decisión óptima.
3. Un determinado método de análisis produce resultados normales $N(x|\mu, \sigma)$. Es necesario estimar un valor para μ y la función de utilidad resulta ser de la forma

$$u(\hat{\mu}, \mu) = 9 - 5(\hat{\mu} - \mu) - (\hat{\mu} - \mu)^2$$

Determinar el estimador óptimo que se deduce de 20 observaciones con valor medio 8,5 y desviación típica 1,4.

4. Determinar la prima máxima que una persona cuya función de utilidad monetaria, en miles de pesetas, es $u(x) = 1 - \exp\{-0,003x\}$, debería pagar para cubrir el riesgo de perder un objeto valorado en 25.000 pesetas, si asigna a ese suceso una probabilidad 0,1 y dispone de un capital total de 200.000 pesetas.
5. Determinar el valor de la información proporcionada sobre μ por 10 observaciones del modelo $N(x|\mu, \sigma)$ con media 15 y desviación típica 3, si se pretende estimar μ , la función de utilidad es $u(\hat{\mu}, \mu) = 10 - (\hat{\mu} - \mu)^2$, y la distribución inicial $N(\mu|12, 3)$.
6. La esperanza de vida de un determinado paciente según se le administre el tratamiento μ_1 o el μ_2 y padezca las enfermedades θ_1, θ_2 o θ_3 viene dada por la tabla (en años esperados de vida).

	θ_1	θ_2	θ_3
μ_1	20	10	5
μ_2	8	10	22

Determinar, en años de vida, el valor que tendría para el paciente conocer la enfermedad que padece si la información de que dispone puede ser descrita por la distribución $p(\theta) = (0,3, 0,3, 0,4)$.

7. La información inicial de que dispone sobre la media μ de la cantidad en μg de la tirosina liberada por ml de plasma sanguíneo en pacientes jóvenes que sufren una úlcera duodenal, puede describirse mediante la distribución $N(\mu|833, 50)$. El valor que tendría para un laboratorio farmacéutico saber si el verdadero valor de μ en estos casos es superior a 750 es de 100.000 ptas. El coste de analizar el plasma

de un conjunto de pacientes es de 20.000 pesetas fijas más mil pesetas por paciente analizado. Determinar el número de pacientes que le resultaría óptimo analizar al laboratorio, si los resultados de dichos análisis se suponen normales con $\sigma = 45$.

8. Se sospecha que existen diferencias significativas entre los resultados obtenidos por dos grupos de alumnos en un examen parcial. Las calificaciones de los 58 alumnos del grupo A tienen una media aritmética de 5,3 y una desviación típica 1,5, mientras que las de los 65 alumnos del grupo B tienen una media de 6,2 y una desviación típica 1,8. Si puede suponerse que las calificaciones de cada grupo constituyen una muestra de una población normal, determinar la probabilidad de que el grupo B sea realmente mejor que el grupo A.
9. Para diagnosticar cual de las dos causas que se consideran posibles, $\theta_1 =$ adenoma o $\theta_2 =$ carcinoma, produce el síndrome de Cushing a un determinado paciente, se analiza el contenido en tetrahidrocortisona de la orina del paciente, obteniéndose 13 mg/24 h. Se dispone además de un banco de datos en el que constan los análisis correspondientes a 6 pacientes con adenoma (3,1, 3,0, 1,9, 3,8, 4,1 y 1,9 mg/24 h) y los correspondientes a 5 pacientes con carcinoma (10,2, 9,2, 9,6, 53,8 y 15,8 mg/24 h). Se sabe que la distribución de los *logaritmos* de las cantidades de tetrahidrocortisona contenidas en la orina tienen una distribución aproximadamente normal. Determinar la distribución diagnóstica del nuevo paciente, si la opinión inicial del equipo médico, vista su historia clínica, es que tiene un carcinoma con probabilidad 0,7.
10. Si, en el problema anterior, se dispone de dos tratamientos μ_1 y μ_2 , la utilidad de cuyas consecuencias puede medirse por la función de utilidad (en años esperados de vida)

	θ_1	θ_2
μ_1	20	2
μ_2	16	10

Determinar el tratamiento óptimo y el valor en años esperados de vida, que después de realizado el análisis de orina y extraídas sus consecuencias, tendría todavía para el paciente conocer la verdadera causa del síndrome que padece.

Referencias

- AITCHISON, J. (1970). *Choice against Chance: An Introduction to Statistical Decision Theory*. Reading, Mass.: Addison-Wesley.
- AITCHISON, J. & DUNSMORE, I.R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.
- ANDERSON, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. Nueva York: Wiley.
- ANSCOMBE, F.J. & AUMANN, R.J. (1963). A definition of subjective probability. *Ann. Math. Statist.*, **34**, 199-205.
- ANSCOMBE, F.J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math.*, **16**, 1-91.
- ASH, R.B. (1972). *Real Analysis and Probability*. Nueva York: Academic Press.
- AYKAC, A. & BRUMAT, S. (eds.) (1977). *New Developments in the Applications of Bayesian Methods*. Amsterdam: North-Holland.
- BARNARD, G.A. (1952). The frequency justification of certain sequential tests. *Biometrika*, **39**, 144-150.
- BARNETT, V. (1973). *Comparative Statistical Inference*. Nueva York: Wiley.
- BARRA, J.R. et al. (eds.) (1977). *Recent Developments in Statistics*. Amsterdam: North-Holland.
- BARTHOLOMEW, D.J. (1965). A comparison of some Bayesian and frequentist inferences. *Biometrika*, **52**, 19-36.
- BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. Reeditado en *Biometrika*, **45**, 293-315, con una nota biográfica escrita por G. A. Barnard.
- BERMÚDEZ, J.D. (1981). La elección de tratamiento como problema de decisión predictiva. *Trab. Estadist.* (en prensa).
- BERNARDO, J.M. (1977). Inferences about the ratio of normal means: a bayesian approach to the Fieller-Creasy problem. En *Recent Developments in Statistics* (Barra et al. eds.), 345-350, Amsterdam: North-Holland.
- BERNARDO, J.M. (1978). Métodos bayesianos y diagnosis clínica. *Estadist. Esp.*, **78/79**, 39-56.

- BERNARDO, J.M. (1979 a). Expected utility as expected information. *Ann. Statist.*, **7**, 686-690.
- BERNARDO, J.M. (1979 b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B*, **41**, 113-147 (con discusión).
- BERNARDO, J.M. (1979 c). Comportamiento asintótico de la información proporcionada por un experimento. *Rev. Acad. Ci. Madrid*, **73**, 491-502.
- BERNARDO, J.M. (1980 a). An information-theoretical approach to approximations in statistics. Conferencia invitada. *12th. European Meeting of Statisticians*, Varna (Bulgaria).
- BERNARDO, J.M. (1981). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadíst.* (En prensa.)
- BERNARDO, J.M. & BASULTO, J. (1979). Análisis Bayesiano de un proceso binomial. *Trab. Estadíst.*, **29**, 3-27.
- BERNARDO, J.M., DEGROOT, M.H., LINDLEY, D.V. & SMITH, A.F.M. (eds.) (1980). *Bayesian Statistics*. Valencia: Imprenta Universitaria.
- BERNARDO, J.M. & GIRON, J. (1980). On the foundations of statistics and decision theory. *Trab. Estadíst.* (En prensa.)
- BETAQUE, N.S. & GORRY, A. (1971). Automating judgmental decision making for a serious medical problem. *Management Sci. B*, **17**, 421-434.
- BORCH, K. (1977). The monster in Loch Ness. En *New Developments in the Applications of Bayesian Methods* (Aykaç & Brunat, eds.), 273-278. Amsterdam: North Holland.
- BOX, G.E.P. & TIAO, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass.: Addison-Wesley.
- BRAITHWAITE, R.B. (1953). *Scientific Explanation*. Cambridge: University Press.
- BROW, L.D. (1964). Sufficient statistics in the case of independent random variables. *Ann. Math. Statist.*, **35**, 14, 56.
- CARNAP, R. (1950). *Logical Foundations of Probability*. Londres: Routledge & Kegan.
- CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics*, **25**, 617-657.
- Cox, D.R. & HINKLEY, D.V. (1974). *Theoretical Statistics*. Londres: Chapman & Hall.
- CRAMER, H. (1946). *Mathematical Methods in Statistics*. Princeton: University Press.
- DARMOIS, G. (1935). Sur les lois de probabilité à estimations exhaustives. *C. R. Acad. Sci. Paris*, **260**, 1265-1266.
- DAVID, F.N. & BARTON, D.E. (1962). *Combinatorial Chance*. Londres: Griffin.
- DAVID, A.P. (1970). On the limiting normality of posterior distributions. *Proc. Cambridge Philos. Soc.*, **67**, 625-633.
- DE FINETTI, B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1-68. Reeditado en inglés en *Studies in Subjective Probability* (Kyburg & Smokler, eds.), 93-158. Nueva York: Wiley, 1964.
- DE FINETTI, B. (1970/1975). *Theory of Probability*. Londres: Wiley.
- DE FINETTI, B. (1972). *Probability, Induction and Statistics*. Nueva York: Wiley.
- DEGROOT, M.H. (1970). *Optimal Statistical Decisions*. Nueva York: McGraw-Hill.
- DEGROOT, M.H. (1975). *Probability and Statistics*. Reading, Mass.: Addison-Wesley.
- DOMBAL, F.T. et al. (1972). Computer-aided diagnosis of acute abdominal pain. *Brit. Med. J.*, **2**, 9-13.
- FELLER, W. (1957/1966). *An Introduction to Probability Theory and its Applications* (2 vols.). Nueva York: Wiley.
- FERRÁNDIZ, J.R. (1980). *Tablas Estadísticas*. Valencia: Editorial Universitaria.
- FIENBERG, S.E. & ZELLNER, A. (eds.) (1975). *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North Holland.
- FINE, T.L. (1973). *Theories of probability an examination of Foundations*. Nueva York: Academic Press.
- FISHER, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London A*, **222**, 309-368. Reeditado en Fisher, R.A. *Contributions to Mathematical Statistics*. Nueva York: Wiley (1950).
- FISZ, M. (1963). *Probability Theory and Mathematical Statistics*. Nueva York: Wiley.
- FREEMAN, H. (1963). *Introduction to Statistical Inference*. Reading, Mass.: Addison-Wesley.
- GARCIA-GARCÍA, J. & LÓPEZ-PELLICER, M. (1978). *Matemáticas COU*. Alcoy: Marfil.
- GINSBERG, A.S. (1970). Decision analysis in clinical patient management. *Proc. 2nd Conf. Diagnostic Processes*. Nueva York: Academic Press.
- GINSBERG, A.S. & OFFENSEND, F.L. (1968). An application of decision theory to a medical diagnosis-treatment problem. *IEEE Trans. Systems, Sci. Cybern.* **SSC-4**, 355-362.
- GNEDENKO, B.V. (1962). *The Theory of Probability*. Nueva York: Chelsea.
- GODAMBE, V.P. & SPROTT, D.A. (eds.) (1971). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart & Winston.
- GOOD, I.J. (1950). *Probability and the Weighing of Evidence*. Londres: Griffin.
- GOOD, I.J. (1965). *The Estimation of Probabilities*. Cambridge, Mass.: The M.I.T. Press.
- GOOD, I.J. (1966). A derivation of the probabilistic explication of information. *J. Roy. Statist. Soc. B*, **28**, 579-581.
- GOOD, I.J. (1969). What is the use of a distribution? En *Multivariate Analysis* (Krishnaiah, ed.), **2**, 183-203. Nueva York: Academic Press.
- GOOD, I.J. (1971). The probabilistic explanation of information, evidence, surprise, causality, explanation and utility. En *Foundations of Statistical Inference*. (Godambe & Sprott, eds.), 108-141 (con discusión). Toronto: Holt, Rinehart & Winston.
- GUSTAFSON, D.H. et al. (1969). Subjective probabilities in medical diagnosis. *IEEE Trans. Man-Machine Systems* **MMS-10**, 61-65.
- HALMOS, P.R. & SAVAGE, L.J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.*, **20**, 225-241.
- HARPER, W.L. & HOOKER, C.A. (eds.) (1976). *Foundations of Probability Theory Statistical Inference and Statistical Theories of Science* (3 vols.). Dordrecht, Holland, Reidel.
- HARTIGAN, J.A. (1964). Invariant prior distributions. *Ann. Math. Statist.*, **35**, 836-845.
- HAUSDORF, F. (1914). *Grünzüge der Mengenlehre*. Leipzig: Teubner.
- HOGG, R.V. & CRAIG, A.T. (1965). *Introduction to Mathematical Statistics*. Nueva York: MacMillan.
- JAYNES, E.T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics*, **SSC-4**, 227-291.

- JAYNES, E.T. (1980). Marginalization and prior probabilities. En *Studies in Bayesian Statistics* (A. Zellner, ed.). Amsterdam: North-Holland.
- JEFFREYS, H. (1939/1967). *Theory of Probability*. Oxford: University Press.
- JOHNSON, R.A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.*, **38**, 1899-1906.
- JOHNSON, R.A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.*, **41**, 851-864.
- JOHNSON, N.L. & KOTZ, S. (1969-1972). *Distributions in Statistics* (4 vols.). Nueva York: Wiley.
- KENDALL, M.G. & STUART, A. (1938/1977). *The Advanced Theory of Statistics* (3 vols.). Londres: Griffin.
- KEYNES, J.M. (1921/1962). *A Treatise on Probability*. Nueva York: Harper Torchbooks.
- KINGMAN, J.F.C. & TAYLOR, S.J. (1966). *Measure and Probability*. Cambridge: University Press.
- KNILL-JONES *et al.* (1973). Use of sequential Bayesian model in diagnosis of jaundice by computer. *Brit. Med. J.*, **1**, 530-533.
- KOLMOGOROV, A.N. (1963). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
- KOOPMAN, B.O. (1963). On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, **39**, 399-409.
- KOOPMAN, B.O. (1940). The bases of probability. *Bull. Amer. Math. Soc.*, **46**, 763-774. Reeditado en *Studies in Subjective Probability* (Kyburg & Smokler, eds.) (1964). Nueva York: Wiley.
- KRICKBERG, K. (1965). *Probability Theory*. Reading, Mass.: Addison-Wesley.
- KYBURG, H.E. & SMOKLER, H. E. (eds.) (1964). *Studies in Subjective Probability*. Nueva York: Wiley.
- LAPLACE, P.S. (1812/1912). *Theorie analytique des probabilités*. Paris: Courcier. Reeditado en *Oeuvres complètes de Laplace* (1912). Paris: Gauthier-Villars.
- LAVALLE, I.H. (1970). *An Introduction to Probability, Decision and Inference*. Nueva York: Holt, Rinehart & Winston.
- LE CAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *Univ. Calif. Publ. Statist.*, **1**, 277-329.
- LE CAM, L. (1958). Les propriétés asymptotiques des solutions de Bayes. *Publ. Inst. Statist. Univ. Pays.*, **7**, 17-35.
- LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, **41**, 802-828.
- LINDGREN, B.W. (1962). *Statistical Theory*. Nueva York: Mac Millan.
- LINDLEY, D.V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. 4th Berkeley Symp. Math. Statist. Probability*, **1**, 453-468. Berkeley: University of California Press.
- LINDLEY, D.V. (1965). *An Introduction to Probability and Statistics from a Bayesian Viewpoint* (2 vols.). Cambridge: University Press.
- LINDLEY, D.V. (1971 a). *Bayesian Statistics, a Review*. Reg. Cont. Ser. Appl. Math., **2**. Philadelphia: SIAM.
- LINDLEY, D.V. (1971 b). *Making Decisions*. Nueva York: Wiley. Traducido al castellano por J. M. Bernardo, *Principios de la Teoría de la Decisión*. Barcelona: Vicens-Vives (1977).
- LINDLEY, D.V. (1978). The Bayesian approach. *Scand. J. Statist.*, **5**, 1-26.
- LOEVE, M. (1955/1977). *Probability Theory* (2 vols.). Berlin: Springer.
- LUSTED, L.B. (1968). *Introduction to Medical Decision Making*. Springfield, Ill.: Thomas.
- MOOD, A.M., GRAYBILL, F.A. & BOES, D.C. (1963/1974). *Introduction to the Theory of Statistics*. Nueva York: McGraw-Hill.
- MURPHY, A.H. & WINKLER, R.L. (1975). Subjective probability forecasting: Some real world experiments. En *Utility, Probability and Human Decision Making* (Wendt & Vlek, eds.), 177-198. Dordrecht, Holland: Reidel.
- NOVICK, M.R. (1969). Multiparameter Bayesian indifference procedures. I. *Roy. Statist. Soc. B*, **31**, 29-64 (with discussion).
- NOVICK, M.R. & JACKSON, P.H. (1974). *Statistical Methods for Educational and Psychological Research*. Nueva York: McGraw-Hill.
- PAPOULIS, A. (1965). *Probability, Random Variables and Stochastic Processes*. Nueva York: McGraw-Hill.
- PARZEN, E. (1960). *Modern Probability Theory and its Applications*. Nueva York: Wiley.
- PHILLIPS, L.D. (1973). *Bayesian Statistics for Social Scientists*. Londres: Nelson.
- PITMAN, E.J.G. (1936). Sufficient statistics and intrinsic accuracy. *Proc. Cambridge Phil. Soc.*, **32**, 567-579.
- PRATT, J.W., RAIFFA, H. & SCHLAIFER, R. (1964). The foundations of decision under uncertainty: an elementary exposition. *J. Amer. Statist. Assoc.*, **59**, 353-375.
- RAIFFA, H. & SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. Cambridge, Mass.: The MIT Press.
- RAMSEY, F.P. (1926). Truth and probability. Reeditado en *Studies in Subjective Probability*. (Kyburg & Smokler, eds.), 61-92. Nueva York: Wiley, 1964.
- RAO, C.R. (1965). *Linear Statistical Inference and its Applications*. Nueva York: Wiley.
- REICHENBACH, H. (1949). *The Theory of Probability*. Berkeley, Calif.: University of California Press.
- RENYI, A. (1962/1966). *Calcul des probabilités avec un appendice sur la théorie de l'information*. Paris: Dunod.
- ROMATKI, V.K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. Nueva York: Wiley.
- SANJUAN, L.F. (1979). *Una generalización del problema de Behrens-Fisher*. Tesis de Licenciatura: Universidad de Valencia.
- SAVAGE, L.R. (1968). *Statistics: Uncertainty and Behaviour*. Boston: Houghton Mifflin.
- SAVAGE, L.J. (1954). *The Foundations of Statistics*. Nueva York: Wiley.
- SAVAGE, L.J. (1961). The foundations of statistics reconsidered. *Proc. 4th Berkeley Symp. Math. Statist. Probability*, **1**, 575-586.
- SAVAGE, L.J. (1970). Diagnosis and the Bayesian viewpoint. *Proc. 2nd Conf. Diagnostic Process*. Nueva York: Academic Press.
- SAVAGE, L.J. (1971). The elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.*, **66**, 783-801.
- SAVAGE, L.J. *et al.* (1962). *The Foundations of Statistical Inference*. Londres: Methuen.

- SCHMITT, R.C. (1969). *An Elementary Introduction to Bayesian Statistics*. Nueva York: Addison-Wesley.
- SENDRA, M. (1979). *Distribución final de referencia para el problema de Pieller-Creasy*. Tesis de Licenciatura: Universidad de Valencia.
- SMITH, A.F.M. & SPIEGELHALTER, D. (1980). Bayes factors and linear model choice criteria. *J. Roy. Statist. Soc. B*, **42**, 213-220.
- SPEHZZLER, C.S. & STAEL VON HOLSTEIN, C.A.S. (1975). Probability encoding in decision analysis. *Management Sci.*, **22**, 340-358.
- SPIEGEL, R.S. (1961/1977). *Estadística*. México: McGraw-Hill.
- TUKEY, J.W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.
- VILLEGAS, C. (1964). On qualitative probability σ -algebras. *Ann. Math. Statist.*, **35**, 1787-1796.
- VON MISES, R. (1936). *Wahrscheinlichkeit, Statistik und Wahrheit*. Viena: Springer.
- WALD, A. (1950). *Statistical Decision Functions*. Nueva York: Wiley.
- WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *J. Roy. Statist. Soc. B*, **31**, 80-88.
- WELCH, B.L. & PEERS, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc. B*, **25**, 318-329.
- WHITWORTS, W.A. (1901). *Choice and Chance*. Cambridge: Deighton Bell.
- WILKS, S.S. (1962). *Mathematical Statistics*. Nueva York: Wiley.
- WINKLER, R.L. (1972). *Introduction to Bayesian Inference and Decision*. Nueva York: Holt, Rinehart & Winston.
- ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Nueva York: Wiley.
- ZELLNER, A. (1977). Maximal data information prior distributions. En *New Developments in the Applications of Bayesian Methods* (A. Aykac and C. Brumat, eds.), 211-132. Amsterdam: North-Holland.
- ZELLNER, A. (ed.) (1980). *Studies in Bayesian Statistics*. Amsterdam: North-Holland.

Soluciones de los problemas

Capítulo 2

- No; $c = 25$ hace el juego equilibrado.
- Debe comercializarse el nuevo fármaco siempre que $p > 1/3$.
- La crema 2, cuya utilidad esperada es 0,64.
- Los comprendidos en el intervalo $0 \leq p < 0,3$.
- Los comprendidos en el intervalo $0 \leq p < 2/3$.
- Si $p > \frac{18}{35}$ la decisión óptima es no anunciar.
- La estrategia óptima consiste en empezar extrayendo una bola de la urna I; si sale blanca, extraeremos otra bola de la urna I, y si sale roja la extraeremos de la urna II.
- Debe apostarse siempre; en contra de los laboristas si $p < 0,6$ y en contra de los conservadores si $p > 0,6$; si $p = 0,6$, debe apostarse, pero es indiferente hacerlo en uno u otro sentido.
- La producción óptima es la de tres máquinas mensuales, que proporciona un beneficio esperado de 10.

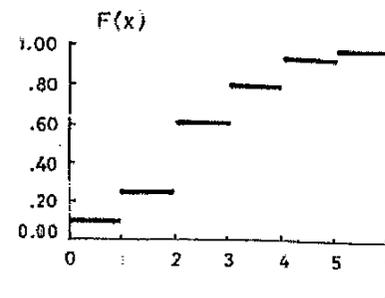
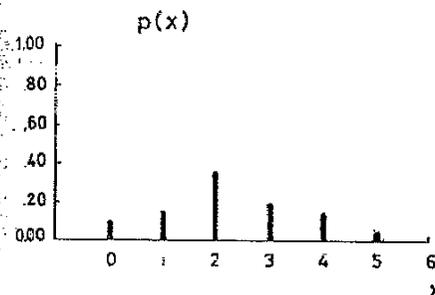
10. El tratamiento óptimo es el t_2 ; la estrategia óptima si los tratamientos pueden ser aplicados consecutivamente consiste en aplicar t_1 y, si no resultase efectivo, aplicar t_2 ; el coste esperado de tal estrategia es 2.

Capítulo 3

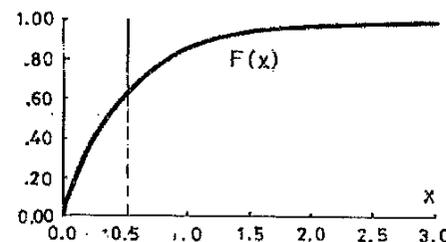
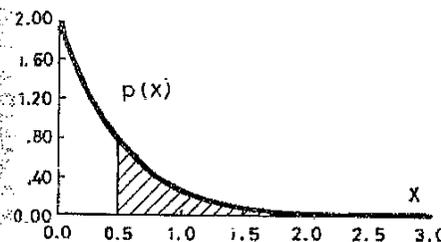
- (Ω, Σ, P) donde $\Omega = \{A, B, C\}$ es el conjunto de los candidatos, $\Sigma = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, \Omega\}$ el álgebra asociada y P la medida sobre ella que asigna respectivamente, las probabilidades $\{0, 0,4, 0,4, 0,2, 0,8, 0,6, 0,6, 1\}$.
- $1 - 0,99^n \geq 0,95$ y por tanto $n \geq 299$.
- En el primer caso, $p(8|10) = 1,861 \times 10^{-5}$, en el segundo, $p(8|10) = (1,861 \times 10^{-5})0,75 + (7,373 \times 10^{-5})0,25 = 3,239 \times 10^{-5}$
- En principio, los sucesos son independientes; caso de existir dependencia sería en sentido contrario: tres hijos varones pueden sugerir una tendencia mayor de lo normal en esa pareja a tener hijos varones, haciendo más alta la probabilidad de un cuarto niño.
- $p(M \cup F) = 3/4$; $1 - p(M \cap F) = 5/6$.
- $p(\text{Alterado}) = 113/360$; $p(\text{Carta } B | \text{Alterado}) = 20/113$.
- Es falso; con una baraja de 40 cartas, sean $A = \{\text{as de espadas}\}$, $B = \{\text{oros}\}$ y $C = \{\text{espadas}\}$ (extracciones sin reemplazamiento).
 $p(B) = 10/40 \vee p(B|A) = 10/39$ luego A favorece a B
 $p(C) = 10/40 \vee p(C|B) = 10/39$ luego B favorece a C
 $p(C) = 10/40 \vee p(C|A) = 9/39$ luego A no favorece a C
- $p(\text{Éxito}) = 0,7$, $p(\text{Fracaso}) = 0,3$.
- Con $p(\text{Sano} | \text{test positivo}) = 0,5$, resulta que $p(\text{Sano}) \approx 0,978$: al estar sana la mayoría de la población, una importante proporción de positivos son falsos positivos. Si $p(\text{Sano})$ fuese, por ejemplo, 0,6, entonces $p(\text{Sano} | \text{test positivo})$ sería tan solo 0,032.
- Para que la probabilidad final sea 0,5 la inicial debería ser 0,918; para que fuese 0,1, tan solo 0,556.

Capítulo 4

- $p[X \geq 2,5] = 0,40$; $p[0 < X < 2] = 0,15$; $p[0 < X \leq 2] = 0,50$; $p[X > 6] = 0$; $p[X = 1,5] = 0$.



- $C = 60/137$; $p[X > 1] = 47/137 \approx 0,343$.
Representaciones gráficas semejantes a las del problema 1.
- $C = 2$; $p[X \geq 0,5] = 1/e \approx 0,368$.



- $p[X < 6] = 0,0198$; $p[6 < X < 7] = 0,768$; $p[6,034 < X < 7,406] = 0,95$.
- $D[X] \approx 5,612$; $p[X > 50] \approx 0,106$.
- $p[\text{Duración} > 4 \text{ horas}] \approx 0,367$.
- $p[X > 0,5] \approx 0,756$.
- | $X - Y$ | p |
|---------|------|
| -3 | 0,02 |
| -1 | 0,22 |
| 1 | 0,49 |
| 3 | 0,27 |

 $E[X - Y] = 1$
 $D[X - Y] \approx 1,51$
- $p[X_2 > 66] \approx 0,251$
 $p[20 < X_1 < 22] = 0,472$
 $p[X_2 > 66 | X = 21] \approx 0,0311$

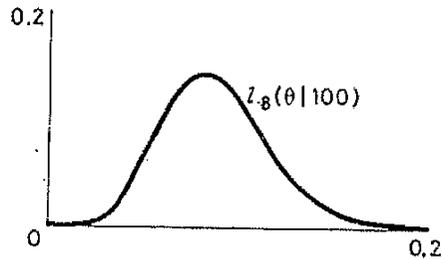
10. $p[X_1 > 800] \approx 0,324$
 $p[X_1 > 800|X_2 = 500] \approx 0,933$

Capítulo 5

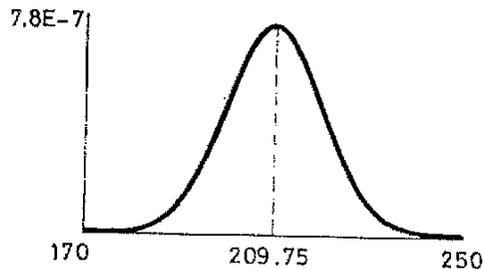
1. $p(0) \approx N(0|22,75, 8,60)$.
 2. Buscando en las tablas de la Beta obtenemos $Be(\theta|8, 8)$.
 Mediante la aproximación a la Normal se obtiene $Be(\theta|7,54, 7,54)$.

3. $I_8(\theta|100) = \binom{100}{8} \theta^8 (1-\theta)^{92}$

Su representación gráfica es:



4. $p(x_1, x_2, x_3, x_4) = \left(\frac{1}{2\pi \cdot 100}\right)^2 \exp\left\{-\frac{1}{2} \sum_{i=1}^4 \left(\frac{x_i - \mu}{10}\right)^2\right\}$

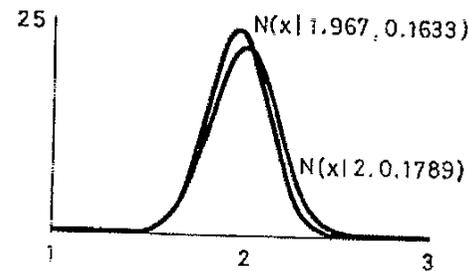


5. $p(x) = N(x|22,75, 9,11)$.
 6. $p(\mu|x) = N(\mu|208,88, 4,772)$.
 7. $p(\theta|x) = Be(\theta|11,5, 123,5)$.

8. $p(\mu|x) = Sx(\mu|1,13, 0,067, 5)$.
 9. $p(x|x_1, x_2) = N(x|24,98, 3,57)$.
 10. Dado que $p(\theta|x) = N(\theta|741,39, 1,35)$ el tratamiento óptimo es \hat{x}_2 , con utilidad esperada 10,92 años.

Capítulo 6

1. $p(\theta|z) = Ga(\theta|5, 12,4)$; $I_{0,95}(\theta) = [0,152, 0,877]$.
 2. $p(x=0) = \int p(x=0|\theta)p(\theta|z)d\theta = \int e^{-3\theta} Ga(\theta|38, 12)d\theta = 0,00021$.
 3. $Gal(x|5, 1)$.
 4. $p(\theta|z) = Be(\theta|142, 130)$; $p(\theta > 0,5|z) \approx 0,767$.
 5. $p(\theta|z) = Be(\theta|1856, 76)$; $I_{0,95}(\theta) \approx [0,953, 0,968]$;
 $I_{0,99}(\theta) \approx [0,948, 0,971]$; $I_{0,999}(\theta) \approx [0,944, 0,973]$.
 6. $p^*(\sigma|s) = N(\sigma|2,2, 0,025)$; $p(\sigma > 2,25|s) = 0,021$.
 7. $\pi(\theta_1, \theta_2) = (2/3, 1/3)$
 $\pi(\theta_1, \theta_2|z) = (0,727, 0,273)$.
 8. $\pi(\theta|x) = N(\theta|2, 0,1789)$, $0 < \theta < 3$.
 9. $p(\theta|x) = N(\theta|1,967, 0,1633)$.



	$\pi(\theta x)$	$p(\theta x)$
$I_{0,9}$	[1,7057, 2,2943]	[1,6984, 2,2356]
$I_{0,99}$	[1,5392, 2,4608]	[1,5464, 2,3876]
$I_{0,999}$	[1,4113, 2,5887]	[1,4297, 2,5043]

10.	$p(z)$	$\pi(z x)$	$p(z x)$
Distr.	$N(z 50, 12,1591)$	$N(z 48,5, 4,4272)$	$N(z 48,6756, 4,1600)$
$I_{0,9}$	[29,999, 70,001]	[41,218, 55,782]	[41,833, 55,518]
$I_{0,99}$	[18,681, 81,319]	[37,096, 59,904]	[37,960, 59,391]
$I_{0,999}$	[9,990, 90,010]	[33,932, 63,068]	[34,987, 62,324]

Capítulo 7

- Rechazamos la hipótesis nula, H_0 , puesto que $p_0 = p(\mu > 23|\bar{x}, s) \approx 0,42 < 0,5$
- Aceptamos la hipótesis nula $H_0: \theta > 0,95$, si $p(H_0) > 0,417$.
Se ajusta $p(\theta) = Be(\theta|85,75, 1,75)$, $p(\theta|x) = Be(\theta|247,75, 9,75)$.
Como $p(H_0|x) = 0,851 > 0,417$ la decisión óptima será aceptar la hipótesis nula.
- El estimador óptimo es $\hat{\mu} = \{2E(\mu|\bar{x}, s) + 5\}/2 = 11$.
- La prima máxima sería 2.586 ptas.
- El valor de la información proporcionada por el experimento es:
 $\nabla u^*(\varepsilon, z) \approx 8,18$.
- El valor de la información perfecta en ese momento es:
 $\nabla u^*(\varepsilon_\infty) = 3,6$ años de vida.
- $p(\mu > 750) = 0,9515$, $H(p) = 0,1941$, $g = 515.603$ y, por tanto, el tamaño muestral óptimo es $n^* = 257$.
- $\theta = \mu_A - \mu_B$, $p(\theta|z) = St(\theta| -0,9, 0,301, 117)$, $p(\theta < 0) \approx 0,9986$.
- $p(\log x|\theta_1) = St(\log x|m_1, s_1\sqrt{\{(m_1 + 1)/(m_1 - 1)\}}}$, $m_1 = 1)$
 $p(\log x|\theta_1) = St(\log x|1,04, 0,359, 5)$
 $p(\log x|\theta_2) = St(\log x|2,71, 0,816, 4)$
 $p(\theta|x) \propto p(\log x|\theta)p(\theta) = (0,01, 0,99)$
- El tratamiento óptimo es t_2 , con una esperanza de vida de 10,06 años. El valor de la información perfecta después de realizado el análisis es tan sólo de 0,04 años ≈ 15 días.

Índice de conceptos

- s-aditividad, 123.
- alcohol, prueba del, 63.
- álgebra, 46
 - engendrada, 46.
- análisis
 - discriminante, 234
 - secuencial, 63.
- aprendizaje, proceso de, 57, 127.
- aproximación
 - de distribuciones, 179
 - normal, 167, 178, 180
 - Poisson de una binomial, 179.
- axiomático, 9
- fundamento, 37.
- Bayes
 - criterio de, 21
 - decisión, 21
 - teorema de, 5, 45, 60, 63, 127, 145, 146.
- Bernouilli
 - sucesiones de, 54, 57, 85
 - sucesos de, 136.
- bits, 148.
- calibrado, 7, 232.
- calidad, control de, 212.
- cantidad
 - aleatoria continua, 88
 - aleatoria X discreta, 83
 - de tirosina, 129, 130, 144, 154, 163, 172.
- cantidades aleatorias, 7, 79, 80, 123, 128.
- características, 103, 109, 118, 167, 202.
- clase exponencial, 175, 203.
- coherencia, principios de, 3, 10, 22, 37, 237.
- combinatoria, 75.
- comparabilidad, 33.
- comportamiento
 - asintótico, 167, 203
 - coherente, 9, 90.
- confianza, nivel de, 169.
- consecuencias, 12.
- correlación, coeficiente de, 119.
- cota
 - inferior, 131
 - superior, 131.
- creencia, grado de, 17, 28, 31, 45.
- criterio condicional, 39.
- coste, 225.
- cuadrado unidad, 26.
- cuantil, 104, 131.
- decisión
 - árbol de, 12
 - colectiva, 36
 - criterios de, 11, 34, 36
 - inadmisible, 21
 - nodo de, 12
 - problema de, 212
 - problemas médicos de, 7, 232
 - proceso lógico de, 11

- teoría de ia. 1, 3, 6, 10, 235
unipersonal, 36.
- decisiones
 espacio de, 10
 sucesivas, 6, 7.
- desviación típica, 5, 103, 168.
- diagnosis, 62, 128, 134, 140, 148, 156, 160, 191.
- diagnóstico, 7, 234.
- discrepancia, 200.
- distancia, 146.
- distribución
 asintótica, 188, 229
 Beta, 91, 99, 141, 174
 normalización de una, 99, 109
 binomial, 85
 de Bernouilli, 85
 de Poisson, 85
 diagnóstica, 235
 exponencial, 92
 final, 127, 145, 146, 174, 178, 200, 212, 237
 final conjunta, 155
 final de referencia, 167, 190, 191, 198, 200
 gamma, 92
 inicial, 127, 128, 137, 143, 145, 174, 200, 203
 conjunta, 155
 normal, 94, 120, 121, 174
 estándar, 95
 multivariante, 120
 predictiva, 139, 143, 144, 146
 prognóstica, 236
 transformada integral de, 100
 uniforme, 89
 χ^2 , 92.
- distribuciones
 admisibles, 192
 continuas, 87
 de probabilidad, 123
 discretas, 83
 gamma
 ajuste de, 182
 cálculo de probabilidades en, 182
 iniciales admisibles, 190
 marginales, 113
 predictivas finales, 161
- Student, 96, 97.
- dominancia, 25.
- enfermedad
 rara, incidencia de una, 194
 tipos de, 114.
- entropía, 147.
- equipos de guardia, 53.
- escuela bayesiana, 4, 75.
- espacio probabilístico, 45, 47.
- esperanza, 102.
- estabilidad, 167.
- estadística
 descriptiva, 5
 media, 5.
- estadístico
 asintóticamente suficiente, 188
 suficiente, 154, 174, 175, 176, 203.
- estimación puntual, 7, 207, 210, 236.
- estimador máximo-verosimil, 137, 139.
- estrategia, 69.
- estudio, forma de, 40.
- exámenes, calificación de, 72.
- experimento
 e, 199
 vacío, 226
 valor esperado del, 226.
- experimentos, diseño de, 6, 7, 224.
- exploraciones peligrosas, 277.
- factorial, función, 2, 52.
- fármaco, comercialización de un, 209.
- familias conjugadas, 173, 174, 176, 177, 203.
- frecuencia relativa, 28.
- función
 característica, 110, 124
 de densidad de probabilidad, 88, 89
 de distribución, 81
 de distribución de un sector aleatorio, 122
 de distribución empírica, 100
 de distribución normal, 2, 79
 de pago, 71
 de probabilidad, 83, 84, 113
 de utilidad, 18, 212, 214
 de utilidad local, 213
 de utilidad propia, 214

- de verosimilitud, 124, 134, 135, 145, 175
- digamma, 182
- gamma, 81
- generatriz, 110
- medible, 97.
- funciones generatrices, 79, 124.
- rundamentos, 4, 7, 9.
- gravedad, centro de, 131.
- hipotesis, contraste de, 7, 207, 236.
- histogramas, 5.
- incertidumbre, ambiente de, 11.
- independencia, 45, 114, 115.
- inducción inversa, 237.
- inferencia estadística, 7, 127.
- infinito numerable, 83.
- información, 146, 149
 cantidad de, 154, 237
 desconocida, cantidad de, 190
 esperada, 147
 inicial, 6
 medida de, 207, 212
 necesaria, 147
 proporcionada, 14
 valor de ia, 66, 207, 219, 220
 perfecta, 221.
- integración numérica, 177, 178.
- intercambiabilidad, 45.
- intercambiables, 56, 124.
- intervalo
 de confianza, 171
 intercuantílico, 104, 105.
- ley
 aditiva, 49
 multiplicativa, 49.
- libertad, grados de, 92.
- límite, teorema central del, 95, 137, 188.
- linealidad, 103, 217.
- líquido cefalorraquídeo, composición del, 189.
- logaritmos neperianos, 99.
- lotería, participación en una, 20.
- magnitud de interés, 215
- matriz
 definida, positiva, 120
 de varianza-covarianza, 119.
- media, 102, 103, 119, 131, 168, 211
 ponderada, 154.
- mediana, 104, 105, 211.
- medida
 de dispersión, 103
 de localización, 103
 descriptiva, 167
 teoría de ia, 103, 123
 unidad de, 26.
- metodología bayesana, 5
 estadística, i, 3.
- métodos
 aproximados, 7, 107
 bayesianos, 9, 42
 estadísticos clásicos, 4.
- moda, 105, 168.
- modelo probabilístico, 145, 199.
- momentos, 79
 absolutos, 103
 centrales, 103.
- Montecarlo, método de, 102.
- morbilidad, tasas de, 132, 142, 150, 151, 161, 168.
- muestra aleatoria, 136.
- muestras grandes, 203.
- muestreo, resultado de un, 200.
- modo aleatorio, 12.
- normal estándar, 179.
- normalidad, test de, 100.
- números aleatorios, 102
 combinatorios, 52.
- observaciones normales, simulación de, 102.
- opción, 23, 32.
- opciones económicas alternativas, 26, 35.
- operación, oportunidad de una, 13.
- oportunidad, pérdida de, 38, 209.
- orina, composición de ia, 196.
- pacientes
 elección de, 84, 105

- hospitalizados, 97.
 Parámetro
 de interés, 128, 145, 159, 199
 marginal, 155
 perturbador, 155.
 partido, pronóstico de un, 70.
 pérdida cuadrática, 210, 211.
 pesos, 95
 y estaturas, 122.
 postulados, 48.
 precisión, 153.
 predicciones, 139, 159.
 premio literario, 59.
 prima probabilística, 217.
 probabilidad, 4, 16, 17, 28, 29, 37
 de cáncer, 87
 de contagio, 186
 definición de, 28, 48
 densidades de, 79, 89
 medida de, 7
 «objetiva», 31
 teorema de la, 58
 teoría de la, 5, 7, 75
 total, 45.
 probabilidades, 4
 absolutas, 31
 finales, 5, 60, 62
 iniciales, 5, 60, 62
 «subjetivas», 42.
 puntos aleatorios, 26, 27.
 recta real, 80.
 referencia, 26
 conjunto de, 216
 consecuencia de, 216
 distribución de, 203.
 regiones creíbles, 169, 202.
 resultados experimentales, 145.
 riesgo
 aversión al, 217
 aversión local al, 218
 ruleta, 54.
 saliva, pH de la, 158
 sangre
 análisis de, 233
 número de hematíes en la, 47
 sensibilidad, análisis de, 198.
 sexo de un recién nacido, 56, 85.
 simetría, 28.
 simulación, 102.
 solución inadmisibles, 41.
status quo, 70.
 sucesos, 45
 ciertos, 16, 23
 inciertos, 12.
 independientes, 53
 sustitución, 25, 26.
 tabla de decisión, 14.
 tamaño
 muestral óptimo, 230
 óptimo de una encuesta, 230.
 temperaturas, 32, 81, 82.
 teoremas básicos, 48.
 teoría normativa, 4.
 test, valor diagnóstico de un, 221.
 tiempos de espera, 106, 111, 177.
 transformación normalizadora, 181,
 182.
 transitividad, 35.
 transporte, elección de un medio de,
 14, 21.
 tratamiento, elección de, 7, 235.
 truncamiento, 195.
 tuberculina, test de, 61.
 tumores, aparición de, 57.
 úlcera, longitud de una, 92.
 utilidad, 4, 18, 32
 definición de, 33
 del dinero, 237
 esperada, 19, 21, 34
 esperada, maximización de la, 36
 garantizada, 38.
 terminal, 225.
 utilidades
 evolución de, 7
 monetarias, 219.
 valor esperado, 118.
 variable aleatoria, 79
 variación, coeficiente de, 103.
 varianzas, 103.
 vectores aleatorios, 79, 112.
 verosimilitud, 4.

Índice de autores

- Aitchison, 235, 237.
 Anderson, 123.
 Anscombe, 42, 203.
 Ash, 123.
 Aumann, 42.
 Aykac, 164.
 Barnett, 164.
 Barton, 75.
 Barra, 164.
 Basuito, 75.
 Bayes, 42, 203.
 Barnard, 203.
 Bartholomew, 204.
 Behrens, 233.
 Bermudez, 236.
 Bernardo, 42, 75, 101, 147, 164, 179,
 181, 191, 203, 215, 229, 234.
 Bêtaque, 237.
 Boes, 123.
 Borca, 31.
 Box, 163, 203.
 Braithwaite, 75.
 Brown, 175.
 Brumat, 164.
 Carnap, 75.
 Cornfield, 42.
 Cox, 164.
 Cragg, 123.
 Cramer, 123.
 David, 75, 203.
 De Finetti, 42, 75, 79, 123, 163.
 DeGroot, 42, 75, 163, 164, 203, 237.
 Dombal, 235.
 Dunsmore, 235, 236.
 Feller, 75, 123.
 Ferrándiz, 2.
 Fienberg, 164.
 Fine, 75.
 Fisher, 203.
 Fisz, 123.
 Freeman, 123.
 García, García, 3.
 Girón, 42.
 Gnedenko, 75, 123.
 Godambe, 164.
 Good, 71, 75, 147, 163, 203.
 Gorry, 237.
 Gravbill, 123.
 Halmos, 203.
 Harper, 164.
 Hartigan, 203.
 Hausdorf, 123.
 Hinkley, 164.

- Hogg, 123.
 Hooker, 164.
- Jackson, 163.
 Jaynes, 203.
 Jettreys, 75, 123, 163, 203.
 Johnson, 123, 183, 203.
 Jones, 235.
- Kendall, 175, 203.
 Keynes, 75.
 Kingman, 123.
 Krill, 235.
 Kolmogorov, 73, 123.
 Koopman, 42.
 Kotz, 123.
 Kriskberg, 123.
 Kvborg, 42, 75.
- Laplace, 42, 75, 123, 203.
 Lavallo, 237.
 Lindgren, 123.
 Lindley, 37, 42, 75, 123, 163, 203,
 204, 218, 237.
 Loeve, 123.
 López-Pellicer, 3.
 Lusted, 237.
- Mood, 123.
 Murphy, 75.
- Novik, 163, 203.
- Papoulis, 123.
 Parzen, 123.
 Peers, 204.
 Phillips, 163.
 Pratt, 42, 218.
- Rao, 123.
 Raiffa, 42, 109, 121, 123, 203, 237.
 Ramsey, 9, 42.
 Reichenbach, 75.
 Rényi, 75, 123, 148.
 Rohatgi, 123.
- Sanjuán, 233.
 Savage, 42, 75, 123, 203.
 Schlaifer, 42, 109, 121, 123, 203, 237.
 Schmitt, 163.
 Sendra, 234.
 Smith, 204.
 Smokler, 42, 75.
 Spetzler, 75.
 Spiegel, 5.
 Spiegelhalter, 204.
 Spratt, 164.
 Staël von Holstein, 75.
 Stuart, 175, 203.
- Taylor, 123.
 Tiao, 163, 203.
 Tukey, 5.
- Villegas, 42.
 Von Mises, 123.
- Wald, 42.
 Wald, 42.
 Walker, 203.
 Walsh, 204.
 Whitworth, 75.
 Wilks, 123.
 Winkler, 75, 163, 237.
- Zellner, 163, 164, 203.

Índice de teoremas

Teorema 2.4.1	29	Teorema 4.6.5	121
Teorema 2.5.1	34	Teorema 4.6.6	121
Teorema 3.2.1	49	Teorema 4.6.7	122
Teorema 3.2.2	50	Teorema 5.4.1	146
Teorema 3.2.3	50	Teorema 5.4.2	147
Teorema 3.2.4	51	Teorema 6.1.1	169
Teorema 3.3.1	54	Teorema 6.2.1	175
Teorema 3.4.1	58	Teorema 6.2.2	176
Teorema 3.4.2	60	Teorema 6.3.1	180
Teorema 4.1.1	81	Teorema 6.3.2	181
Teorema 4.2.1	86	Teorema 6.3.3	182
Teorema 4.3.1	89	Teorema 6.4.2	187
Teorema 4.4.1	98	Teorema 6.4.3	188
Teorema 4.4.2	100	Teorema 6.5.1	191
Teorema 4.4.3	101	Teorema 6.5.2	192
Teorema 4.5.1	103	Teorema 6.5.3	193
Teorema 4.5.2	104	Teorema 6.5.4	198
Teorema 4.5.4	109	Teorema 7.1.1	211
Teorema 4.5.5	110	Teorema 7.1.2	211
Teorema 4.6.2	119	Teorema 7.2.1	215
Teorema 4.6.3	119	Teorema 7.5.1	229
Teorema 4.6.4	120		

Índice de ecuaciones

Ecuación 2.2.1	18	Ecuación 3.5.4	63
Ecuación 2.2.2	19	Ecuación 3.5.5	63
Ecuación 2.2.3	21	Ecuación 3.6.1	70
Ecuación 2.2.4	21	Ecuación 3.6.2	70
Ecuación 2.3.1	23	Ecuación 3.6.3	70
Ecuación 2.4.1	32	Ecuación 3.6.4	71
Ecuación 2.5.1	33	Ecuación 3.6.5	72
Ecuación 2.5.2	33	Ecuación 3.6.6	73
Ecuación 2.5.3	33	Ecuación 4.2.1	87
Ecuación 2.5.4	33	Ecuación 4.3.1	94
Ecuación 2.5.5	34	Ecuación 4.3.2	94
Ecuación 2.5.6	34	Ecuación 4.3.3	95
Ecuación 2.5.7	34	Ecuación 4.4.1	97
Ecuación 2.5.8	34	Ecuación 4.6.1	113
Ecuación 2.5.9	35	Ecuación 4.6.2	113
Ecuación 2.5.10	35	Ecuación 4.6.3	113
Ecuación 3.2.1	50	Ecuación 4.6.4	121
Ecuación 3.2.2	51	Ecuación 5.1.1	131
Ecuación 3.2.3	52	Ecuación 5.1.2	131
Ecuación 3.3.1	53	Ecuación 5.1.3	131
Ecuación 3.3.2	53	Ecuación 5.1.4	132
Ecuación 3.3.3	54	Ecuación 5.1.5	132
Ecuación 3.3.4	54	Ecuación 5.1.6	132
Ecuación 3.3.5	55	Ecuación 5.1.7	132
Ecuación 3.4.1	60	Ecuación 5.2.1	136
Ecuación 3.4.2	60	Ecuación 5.2.2	139
Ecuación 3.4.3	61	Ecuación 5.3.1	139
Ecuación 3.5.1	63	Ecuación 5.3.2	139
Ecuación 3.5.2	63	Ecuación 5.3.3	141
Ecuación 3.5.3	63	Ecuación 5.3.4	141

Ecuación 5.3.5	141	Ecuación 6.1.1	171
Ecuación 5.3.6	142	Ecuación 6.1.2	171
Ecuación 5.3.7	143	Ecuación 6.1.3	171
Ecuación 5.3.8	143	Ecuación 6.1.4	172
Ecuación 5.4.1	146	Ecuación 6.1.5	173
Ecuación 5.4.2	146	Ecuación 6.2.1	175
Ecuación 5.4.3	146	Ecuación 6.2.2	175
Ecuación 5.4.4	146	Ecuación 6.3.1	181
Ecuación 5.4.5	147	Ecuación 6.4.1	184
Ecuación 5.4.6	147	Ecuación 6.4.2	184
Ecuación 5.4.7	147	Ecuación 6.4.3	185
Ecuación 5.4.8	148	Ecuación 6.4.4	185
Ecuación 5.4.9	148	Ecuación 6.4.5	185
Ecuación 5.4.10	150	Ecuación 6.4.6	185
Ecuación 5.4.11	151	Ecuación 6.4.7	185
Ecuación 5.4.12	152	Ecuación 6.4.8	189
Ecuación 5.4.13	153	Ecuación 6.4.9	189
Ecuación 5.4.14	153	Ecuación 6.5.1	193
Ecuación 5.4.15	153	Ecuación 6.5.2	195
Ecuación 5.4.16	153	Ecuación 6.5.3	196
Ecuación 5.4.17	153	Ecuación 7.1.1	210
Ecuación 5.4.18	153	Ecuación 7.1.2	211
Ecuación 5.4.19	153	Ecuación 7.1.3	211
Ecuación 5.4.20	153	Ecuación 7.2.1	215
Ecuación 5.4.21	154	Ecuación 7.3.1	218
Ecuación 5.5.1	157	Ecuación 7.4.1	220
Ecuación 5.5.2	157	Ecuación 7.4.2	220
Ecuación 5.5.3	157	Ecuación 7.4.3	220
Ecuación 5.5.4	157	Ecuación 7.4.4	220
Ecuación 5.5.5	157	Ecuación 7.4.5	223
Ecuación 5.5.6	158	Ecuación 7.5.1	225
Ecuación 5.6.1	159	Ecuación 7.5.2	230
Ecuación 5.6.2	159	Ecuación 7.5.3	231
Ecuación 5.6.3	161	Ecuación 7.5.4	231
Ecuación 5.6.4	162	Ecuación 7.6.1	233

Índice de definiciones

Definición 2.4.1	29	Definición 4.5.3	104
Definición 2.5.1	33	Definición 4.5.4	105
Definición 3.1.1	46	Definición 4.5.5	110
Definición 3.1.2	47	Definición 4.6.1	112
Definición 3.3.1	53	Definición 4.6.2	112
Definición 3.3.2	56	Definición 4.6.3	113
Definición 4.1.1	80	Definición 4.6.4	115
Definición 4.1.2	81	Definición 4.6.5	115
Definición 4.2.1	83	Definición 4.6.6	115
Definición 4.2.2	85	Definición 4.6.7	120
Definición 4.2.3	85	Definición 5.2.1	136
Definición 4.2.4	87	Definición 5.2.2	137
Definición 4.3.1	88	Definición 5.3.1	141
Definición 4.3.2	89	Definición 5.4.1	147
Definición 4.3.3	91	Definición 5.4.2	147
Definición 4.3.4	92	Definición 6.1.1	169
Definición 4.3.5	94	Definición 6.2.1	174
Definición 4.3.6	97	Definición 6.2.2	174
Definición 4.5.1	102	Definición 6.5.1	190
Definición 4.5.2	103		

Índice de ejemplos

Ejemplo 2.1.1	13	Ejemplo 4.4.3	100
Ejemplo 2.1.2	14	Ejemplo 4.4.4	102
Ejemplo 2.2.1	20	Ejemplo 4.5.1	105
Ejemplo 2.2.2	21	Ejemplo 4.5.2	106
Ejemplo 2.3.1	27	Ejemplo 4.5.3	109
Ejemplo 2.4.1	32	Ejemplo 4.5.4	111
Ejemplo 2.5.1	35	Ejemplo 4.6.1	114
Ejemplo 2.6.1	41	Ejemplo 4.6.2	115
Ejemplo 3.1.1	47	Ejemplo 4.6.3	117
Ejemplo 3.2.1	51	Ejemplo 4.6.4	122
Ejemplo 3.2.2	53	Ejemplo 5.1.1	128
Ejemplo 3.3.1	54	Ejemplo 5.1.2	129
Ejemplo 3.3.2	56	Ejemplo 5.1.3	130
Ejemplo 3.3.3	57	Ejemplo 5.1.4	132
Ejemplo 3.4.1	59	Ejemplo 5.2.1	134
Ejemplo 3.4.2	61	Ejemplo 5.3.1	140
Ejemplo 3.4.3	62	Ejemplo 5.3.2	142
Ejemplo 3.5.1	63	Ejemplo 5.3.3	144
Ejemplo 3.5.2	66	Ejemplo 5.3.4	144
Ejemplo 3.6.1	70	Ejemplo 5.4.1	148
Ejemplo 3.6.2	72	Ejemplo 5.4.2	150
Ejemplo 4.1.1	81	Ejemplo 5.4.3	151
Ejemplo 4.1.2	82	Ejemplo 5.4.4	154
Ejemplo 4.2.1	84	Ejemplo 5.5.1	156
Ejemplo 4.2.2	85	Ejemplo 5.5.2	158
Ejemplo 4.2.3	87	Ejemplo 5.6.1	160
Ejemplo 4.3.1	92	Ejemplo 5.6.2	161
Ejemplo 4.3.2	95	Ejemplo 5.6.3	163
Ejemplo 4.4.1	97	Ejemplo 6.1.1	168
Ejemplo 4.4.2	99	Ejemplo 6.1.2	172

Ejemplo 6.2.1	177	Ejemplo 6.6.1	200
Ejemplo 6.3.1	179	Ejemplo 7.1.1	209
Ejemplo 6.3.2	182	Ejemplo 7.1.2	212
Ejemplo 6.3.3	182	Ejemplo 7.3.1	219
Ejemplo 6.4.1	186	Ejemplo 7.4.1	221
Ejemplo 6.4.2	189	Ejemplo 7.5.1	227
Ejemplo 6.5.1	191	Ejemplo 7.5.2	230
Ejemplo 6.5.2	194	Ejemplo 7.6.1	233
Ejemplo 6.5.3	196		

Índice de símbolos

A, 17, 45	m_x , 103
$Bb(x \alpha, \beta, n)$, 141	$N(x \mu, \sigma)$, 94
$Be(x \alpha, \beta)$, 91	$N(x 0, 1)$, 95
$Bk(x \theta, n)$, 85	$n!$, 52
$Br(x \theta)$, 85	$\binom{n}{r}$, 52
C, 190	$P(A B, H)$, 60
c_x, c^* , 24	$p(x)$, 83, 88
C_{ij} , 12	$p(x \theta)$, 134
d_x , 12	$Po(x \lambda)$, 87
$D[x]$, 103	$p(\theta_i d_i, H)$, 18
$D^2[x]$, 103	$p(A H)$, 17, 29, 45
$D^2[Y]$, 109	$p(A \cup B H)$, 48, 51
E, 225	r , 136
$E[f(x)]$, 102	R, 80
$E(x)$, 103	R_i , 26
$Ex(x \beta)$, 92	s^2 , 139, 153
$E[Y]$, 109	$St(x \mu, \sigma, a)$, 97
$F(x)$, 81	u_0^* , 220
$Ga(x \alpha, \beta)$, 92	$u^*(\epsilon, z, d)$, 220
H, 17, 45	$u^*(\epsilon, z)$, 226
$H\{p(\theta)\}$, 147	$u(c)$, 33
$I^0\{\epsilon, p(\theta) x\}$, 147, 148	$u(i)$, 18
$I_p(\theta)$, 169	$u^*(d_i)$, 34
$l = \{c_1 A_1, c_2 A_2, \dots, c_k A_k\}$, 23	$u\{p_1, \dots, p_k, \theta^*\}$, 71
$i_x(\theta)$, 134	$Un(x \alpha, \beta)$, 90
	x , 134
	\bar{x} , 139, 153

268 BIOESTADISTICA

$\Gamma(x)$, 91
 ε , 134
 $\varepsilon(k)$, 190
 θ_{ij} , 12
 θ , 137
 Θ_i , 12
 μ_k , 103

σ , 60
 $\pi(\theta)$, 190
 $\pi(\theta|x)$, 191
 Σ , 46, 47
 $\Phi(x)$, 95
 χ^2 , 92
 $\Psi(t)$, 110
 Ω , 23, 45, 47