A.A. Borovkov

ESTADÍSTICA *matemática*

Editorial Mir Moscú

A.A. Borovkov

ESTADÍSTICA *matemática*

Estimación de los parámetros
Verificación de las hipotesis
Capítulos adicionales

Editorial Mir Moscú

Traducido del ruso por A. Samojválov
Impreso en la URSS
На испанском языке

ISBN 5—03—000621—4.

Índice

Capítulo 4

Problemas estadísticos de dos muestras y más

Capítulo 5

Enfoque de los problemas de la estadística matemática desde el punto de vista de la teoría de los juegos

Prefacio

Este libro se basa en las conferencias de estadística matemática que el autor dictó durante muchos años en el tercer curso de la facultad de matemáticas de la Universidad de Novosibirsk. Con el andar del tiempo, el curso de conferencias ha sido varias veces modificado en busca de una variante que fuera, en la medida de lo posible, más armoniosa y accesible, y que al mismo tiempo correspondiera al estado moderno de esta ciencia. Se probaron distintas variantes, comenzando por un curso de carácter principalmente prescriptivo, con la exposición de los tipos básicos de problemas (construcción de estimaciones y criterios y estudio de sus propiedades), y terminando por un curso de carácter general, dedicado a la teoría de los juegos, en el que la teoría de las estimaciones y la verificación de las hipótesis eran no más que casos particulares de un enfoque único. A consecuencia del tiempo limitado (un semestre) no fue posible unificar dichas variantes íntimamente ligadas, cada una de las cuales poseía, por separado, defectos evidentes. En el primer caso, el conjunto de hechos concretos obstaculizaba el desarrollo de una opinión general en cuanto al objeto de estudio. La segunda variante carecía de resultados concretos sencillos y estaba sobrecargada de muchos conceptos nuevos, muy complejos, cuya asimilación constituía una tarea extraordinariamente difícil. Por lo visto, la más conveniente es la variante en la que la exposición de los elementos de la teoría de las estimaciones y de la teoría de verificación de las hipótesis concuerda con el mantenimiento consecutivo de la línea de búsqueda de los procedimientos óptimos.

Los capítulos fundamentales del libro se basan en el material unificado de las conferencias impartidas en tiempos diferentes y ampliadas a expensas de los apartados cuya presencia ha sido dictada por la propia lógica de exposición. El objetivo principal consiste en aclarar el estado actual de la

materia en concordancia con su accesibilidad máxima posible y la integridad y armonía matemática.

El libro comprende 5 capítulos y 8 suplementos.

En el capítulo 1 se estudian las propiedades (fundamentalmente asintóticas) de las distribuciones empíricas, que constituyen la base de la estadística matemática.

En los capítulos 2 y 3 se ofrecen, respectivamente, la teoría de las estimaciones y la teoría de verificación de las hipótesis estadísticas. Las primeras partes de cada uno de estos capítulos están dedicadas a la descripción de los posibles enfoques de la resolución de los problemas planteados, así como a la búsqueda de los procedimientos óptimos. Las segundas partes ofrecen la construcción de los procedimientos asintóticamente óptimos.

El capítulo 5 tiene esa misma estructura. En él se expone el enfoque general de los problemas de la estadística matemática desde el punto de vista de la teoría de los juegos.

El capítulo 4 está dedicado a los problemas relacionados con dos muestras y más.

Los suplementos del libro se hallan vinculados a las afirmaciones en el texto principal, cuya demostración sale fuera del marco de la exposición fundamental, ya por su carácter, ya por su dificultad.

El manual también contiene observaciones bibliográficas que no pretenden ser completas, pero que permiten seguir el surgimiento y el desarrollo de las principales tendencias de la estadística matemática. Además, por doquier donde ha sido posible, se ha dado preferencia a las alegaciones monográficas (como el tipo de literatura más accesible) y no a los artículos originales.

Hoy día existen bastantes manuales de estadística matemática. Entre ellos cabe destacar los cuatro siguientes, en cuyas páginas se expone un amplio material que refleja el estado actual de la materia: son los libros de H. Cramer [25], E. Lehmann [57], S. Zacks [95], I.A. Ibraguímov y R.Z. Jasminski [48]. Pero la máxima influencia en la escritura de la obra presente fue ejercida por las monografías [48] (algunas ideas de este libro se han utilizado en los §§ 23—25, 27—29 del cap. 2) y [57] (la exposición de los §§ 5—8 del capítulo 3 se asemeja, por su contenido, a los respectivos apartados de [57]). La demás exposición está poco relacionada, según su estructura, con los libros mencionados.

Hay muchas otras obras que ocupan un lugar notable en la literatura estadística (tales como los libros de Blackwell y Girshak [7], Kendall y Stuart [49, 50], Cox y Hinkly [23], Ferguson [33], Rao [76] y una serie de otros — no hay posibilidad de presentar su enumeración completa), pero por su espíritu y por la selección del material, estos trabajos se distinguen

considerablemente de la monografía que se ofrece a la atención de los lectores *).

A la par con los resultados y enfoques conocidos, en el libro presente se han incluido algunos apartados nuevos que simplifican la exposición del material, se han hecho varias mejoras metodológicas y se han utilizado algunos resultados nuevos, así como resultados que se publican por primera vez en la literatura monográfica.

A continuación se ofrece una descripción breve de la estructura metodológica del libro (véanse también el índice y los prefacios breves de cada uno de los capítulos).

En los §§ 1 y 2 del capítulo 1 se intrducen los conceptos de muestra y de distribución empírica y se establece el teorema de Glivenko — Cantelli, el cual puede considerarse como un hecho fundamental que constituye la base de las deducciones estadísticas.

En § 3 se introducen dos tipos de estadísticas (de los tipos I y II) que comprenden la inmensa mayoría de las estadísticas prácticamente interesantes, las cuales se definen como valores $G(\mathbf{P}_n^*)$ de las funcionales G (que satisfacen ciertas condiciones) de la distribución empírica $\mathbf{P}_n^*$. Más adelante, en los §§ 7 y 8 se establecen los teoremas del límite de distribución de dichas estadísticas. Esto simplifica la exposición posterior y permite no citar, para cada estadística concreta, prácticamente los mismos razonamientos que no se refieren, además, a la esencia de la cuestión.

En el § 5 han sido reunidos los teoremas auxiliares (que en el libro se denominan "teoremas de continuidad") sobre la convergencia de las distribuciones y la convergencia de sus momentos. Esto también simplifica la exposición posterior.

En el § 6 (no obligatorio en la primera lectura del libro) se establece que la función empírica de distribución $F_n^*(t)$ es un proceso poissoniano condicional, y se ofrece la enunciación del teorema (demostrado en el suplemento I) de la convergencia des proceso $\sqrt{n}(F_n^*(t) - F(t))$ hacia el puente browniano.

En el § 10 se introducen las distribuciones empíricas suavizadas que permiten aproximar no sólo la propia distribución, sino también su densidad.

En el § 3 del capítulo 2, dedicado a las estimaciones de los parámetros desconocidos, se introduce un método único de construcción de las estimaciones, denominado "método de sustitución". Este consiste en que la estimación θ^* para el parámetro θ, representado en forma de la funcional $\theta = G(\mathbf{P})$ de la distribución $\mathbf{P}$ de la muestra, es preciso buscarla en forma

*) En el año 1983 apareció un magnífico libro de *E. Lehmann* [58], en el cual, en adición a [57], se expone la actual teoría de estimación.

de $\theta^* = G(\mathbf{P}_n^*)$, donde $\mathbf{P}_n^*$ es la distribución empírica. Todas las estimaciones "razonables" usadas en la práctica son estimaciones de sustitución. La optimación de una estimación se alcanza eligiendo una funcional conveniente G. Si la estadística $\theta^* = G(\mathbf{P}_n^*)$ es de los tipos I ó II, los teoremas del capítulo 1 permiten establecer en seguida la validez de estas estimaciones y su normalidad asintótica. En los §§ 4 y 5, este enfoque es ilustrado por las estimaciones obtenidas mediante el método de momentos y el método de distancia mínima. Desde esas mismas posiciones también se podrían examinar las estimaciones de máxima verosimilitud (§ 6), pero su estudio inmediato da la posibilidad de obtener resultados más profundos, que serán necesarios ulteriormente.

La comparación de las estimaciones del capítulo 2 se realiza a base de dos enfoques: *estándar* o *medio cuadrático* (se comparan $\mathrm{M}_\theta\,(\theta^* - \theta)^2$ y *asintótico* (se comparan las varianzas de la distribución límite $\sqrt{n}(\theta^* - \theta)$ en la clase de estimaciones asintóticamente normales). En el caso paramétrico, esto permite destacar 3 tipos de estimaciones óptimas: estimaciones eficientes en las clases K_b, con un desplazamiento fijo b, y estimaciones bayesianas y minimax. A base de esos mismos principios se separan las clases de estimaciones *asintóticamente* óptimas en el enfoque asintótico. Para construir las estimaciones eficientes se utilizan los siguientes métodos tradicionales: el primero tiene carácter cualitativo y está vinculado al principio de suficiencia (§§ 12—14); el segundo se basa en las relaciones cuantitativas que se deducen de la desigualdad de Rao — Cramer (§ 16); y el tercero se halla relacionado con las consideraciones de invariación (§§ 17 y 19) que permiten reducir la clase de las estimaciones sometidas a examen.

Los §§ 20—30 están dedicados a la determinación de las estimaciones asintóticamente óptimas y al estudio de las propiedades asintóticas de la función de verosimilitud. El párrafo 20 contiene la desigualdad integral del tipo Rao — Cramer que permite, en particular, obtener criterios simples de carácter asintóticamente bayesiano y minimax de las estimaciones, así como fundamentar la separación de cierta subclase de estimaciones K_0 a la cual conviene limitarse en búsqueda de estimaciones asintóticamente eficientes. Esto da la posibilidad de establecer inmediatamente en el § 25, mediante el estudio de las propiedades asintóticas de las estimaciones de verosimilitud máxima, el carácter asintóticamente bayesiano y minimax de las estimaciones mencionadas, así como su eficiencia asintótica en K_0. Los párrafos 21—24 tienen carácter auxiliar. La estimación de los parámetros por intervalos se examina en los §§ 31 y 32 y también en el § 8 del capítulo 3.

El capítulo 3 está dedicado a la verificación de las hipótesis. En los §§ 1 y 2 se examina el caso de un número finito de hipótesis simples. Se

destacan (de un modo análogo a la teoría de estimación) tres tipos de criterios óptimos: los más potentes en sus subclases, los bayesianos y los minimax. Se establecen las relaciones entre estos criterios y se determina su forma evidente. Además, las consideraciones se basan en el principio bayesiano (y no en el lema de Neyman — Pearson) lo que, a nuestro juicio, simplifica la exposición y hace más comprensible el material. En el § 3 se examinan los enfoques asintóticos del cálculo de los criterios para verificar dos hipótesis simples y se realiza su comparación. En el § 4 se analiza el planteamiento general del problema sobre la verificación de dos hipótesis compuestas y se definen las clases de criterios óptimos (uniformemente más potentes, bayesianos y minimax). El párrafo 5 está dedicado a la búsqueda de criterios uniformemente más potentes en los casos cuando esto es posible. En los §§ 6 y 7 se resuelve el mismo problema, pero en las clases de criterios contraídos a base de consideraciones de no desplazamiento y de invariación. Además, al igual que en los §§ 1 y 2, las consideraciones se basan en el enfoque bayesiano. En el § 8 se construyen, con ayuda de los resultados obtenidos, los conjuntos confidenciales más exactos. En el § 9 se examinan los criterios bayesianos y minimax. Los párrafos 10 y 13 están dedicados al criterio de la relación de verosimilitud. Este criterio resulta uniformemente el más potente en muchos casos particulares y posee carácter asintóticamente bayesiano para conjeturas bastante amplias. El estudio de las propiedades de optimación asintótica del criterio de la relación de verosimilitud continúa en los §§ 15—17. En el § 11 se establece el valor óptimo de este criterio en los problemas del análisis sucesivo. Los párrafos 14 y 15 están dedicados a la búsqueda de criterios asintóticamente óptimos para verificar las hipótesis afines, y se ha encontrado su forma explícita simple para los principales problemas estadísticos.

Una particularidad importante de los tres primeros capítulos es el hecho de que en ellos se examinan tan sólo los problemas estadísticos relacionados con la utilización de una muestra.

Como ya fue señalado, el capítulo 4 del libro está dedicado a los problemas de dos muestras y más. A ellos pertenecen, antes que nada, los problemas sobre la homogeneidad (completa o parcial, §§ 1 y 2) y los problemas de regresión (§ 3) y del análisis de varianza (§ 4). A base de los resultados del capítulo 3, para los problemas de homogeneidad (en el caso paramétrico) se han construido los criterios asintóticamente óptimos, suponiendo que las hipótesis alternativas son semejantes a la hipótesis principal sobre la homogeneidad. Para los problemas de regresión (tanto para la regresión lineal como para la relacionada con las funciones arbitrarias) se han hallado, con ayuda de los resultados de los capítulos 2 y 3, las estimaciones eficientes de los parámetros desconocidos y se han construido los criterios para verificar las hipótesis principales. También han sido examinados los

llamados problemas de reconocimiento de imágenes (§ 5), los cuales, por lo visto, aparecen por primera vez en la literatura didáctica.

El capítulo 5 está dedicado al enfoque general de los problemas de estadística desde el punto de vista de la teoría de los juegos. Este enfoque contribuye a la formación de una opinión general acerca del objeto de estudio de la estadística matemática y permite generalizar muchos resultados de los capítulos 2 y 3. En el § 2 se exponen los conceptos y resultados principales de la teoría "ordinaria" de los juegos (se examinan únicamente los juegos de dos personas). En particular, se establecen las relaciones entre los tipos principales de estrategias óptimas: bayesianas, minimax y las uniformemente mejores en las subclases. En el § 3 se estudian los juegos estadísticos. En el § 4 se enuncia y se demuestra el llamado principio bayesiano que permite reducir el problema de búsqueda de la resolución estadística bayesiana a un problema mucho más fácil de construcción de la estrategia bayesiana para el juego ordinario de dos personas. En el § 5 se analizan los principios de suficiencia, de no desplazamiento y de invariación para construir las resoluciones uniformemente mejores en las subclases respectivas. Los párrafos 6—8 están dedicados a la búsqueda de las reglas decisivas asintóticamente óptimas. En el § 6 se estudian las estimaciones asintóticamente óptimas de los parámetros para la función arbitraria (y no sólo cuadrática) de pérdidas. En este caso se logra establecer los resultados semejantes a los del cap. 2 sobre la optimación asintótica de las estimaciones de verosimilitud máxima. En los § 7 y 8 se examinan los criterios asintóticamente óptimos para la función arbitraria de pérdidas. En el § 7 se demuestra el criterio asintóticamente bayesiano de la relacion de verosimilitud; en el § 8 se establece el indicio límite de optimación de los criterios para verificar las hipótesis semejantes (generalización de los resultados de los §§ 14 y 15 del cap. 3 para el caso de una función arbitraria de pérdidas).

Entre los Suplementos cabe destacar el Suplemento VIII donde se demuestran dos teoremas fundamentales de la teoría de los juegos estadísticos y cuya lectura exige una preparación matemática más alta.

El libro tiene muchas finalidades. Claro está que en su volumen completo, el mismo se asemeja más al programa mínimo para el curso de postgraduados de la especialidad de "Estadística Matemática", que a un libro de texto para los estudiantes. Pero en esta obra se prevé un sistema de medidas que facilitan su primera lectura y que la hacen accesible también para los estudiantes. Los párrafos de elevada dificultad o "más avanzados" en cuanto a su contenido están anotados con un asterisco y conviene omitirlos al leerlos por primera vez, así como el texto escrito con letra gallarda. Además, la exposición de los casos técnicamente más complicados, relaciona-

dos con el parámetro multidimensional, casi siempre se ofrece en apartados y párrafos independientes que también pueden ser omitidos.

Los profesores de los centros de enseñanza superior que ya conocen, al menos parcialmente, la asignatura pueden escoger del libro un conjunto de párrafos (puede haber muchas variantes) a base de los cuales (no es obligatorio utilizarlos por completo) es posible componer un curso semestral de estadística matemática. He aquí una de las variantes: §§ 1, 3 y 5 del capítulo 1; §§ 2—4, 6—12, 14, 16, (21, 23—25), 31 y 32 del capítulo 2; §§ 1, 2, 4, 5, 12 (13, 16) del capítulo 3. Los párrafos entre paréntesis están dedicados a los procedimientos asintóticamente óptimos. Según el grado de preparación de los estudiantes, es necesario organizar la enseñanza de dichos párrafos de la forma más accesible u omitirlos por completo.

La lectura del libro supone el conocimiento del curso de la teoría de las probabilidades conforme al volumen del manual de A.A. Borovkov [11]. Las remisiones a este libro, a diferencia de otras, aparecen en los lugares que el lector, por lo visto, debe conocer, y sirven fundamentalmente para hacer memoria.

La numeración de los párrafos en cada capítulo del libro es independiente, así como la de los teoremas (lemas, ejemplos, etc.) en cada párrafo. A fin de hacer más cómoda la lectura se utilizan diversos sistemas para las referencias a los teoremas, lemas, ejemplos, fórmulas, etc., según su alejamiento del pasaje que se lee. Si se hace una referencia al teorema 1 o a la fórmula (12) del párrafo que se lee, la misma se escribirá del siguiente modo: teorema 1, fórmula (12). Si se trata del teorema 1 y la fórmula (12) de uno de los párrafos precedentes de este capítulo (por ejemplo, del § 13), la referencia tendrá la forma siguiente: teorema 13.1, fórmula (13.12). Por último, si se hacen referencias a otro capítulo, aparecerá, además, el indicador del número de este último (primera cifra). Por ejemplo, el teorema 2.13.1 denota el teorema 1 del § 13 del capítulo 2, y la fórmula (2.13.12) denota la fórmula (12) del § 13 del capítulo 2. Eso mismo corresponde a la designación de los párrafos. La referencia al § 13 significa la remisión al § 13 de este capítulo, y la referencia al § 2.13 significa la remisión al § 13 del capítulo 2.

El signo $\triangleleft$ significa la terminación de la demostración.

Para facilitar la lectura del libro, al final de éste se da la lista de las principales designaciones y se expone el índice alfabético de materias.

A.A. Borovkov

Introducción

En el presente libro se exponen los fundamentos de la parte de las matemáticas que se llama *estadística matemática*. Para abreviar, esta última suele denominarse simplemente *estadística*. Sin embargo, conviene tener presente que tal abreviación sólo es posible cuando existe una buena comprensión mutua, puesto que, de por sí, el término "estadística" corresponde generalmente a un concepto algo distinto.

¿Qué representa la asignatura de estadística matemática? Se pueden citar diversas "definiciones" descriptivas que reflejan, en mayor o menor grado, el contenido de esta parte de las matemáticas. Una de las definiciones más simples y aproximadas se basa en la comparación relacionada con el concepto de selección de muestras de la población madre, así como con el problema de distribución hipergeométrica que se examina, por regla general, al principio del curso de teoría de las probabilidades. Conociendo la composición de la población madre, allí se estudian las distribuciones para la composición de una muestra aleatoria. Es un *problema directo* típico de la teoría de las probabilidades. No obstante, frecuentemente también es preciso resolver *problemas recíprocos* cuando se conoce la composición de la muestra y, basándose en ella, es necesario determinar cómo era la población madre. Tales tipos de problemas recíprocos son los que en realidad constituyen, hablando metafóricamente, la asignatura de estadística matemática.

Precisando algo esta comparación se puede decir lo siguiente: en la teoría de las probabilidades, conociendo la naturaleza de cierto fenómeno, aclaramos cómo se comportarán (cómo están distribuidas) unas u otras características sujetas a estudio, que pueden ser observadas en los experimentos. En la estadística matemática sucede al revés: como material de partida sirven los datos experimentales (generalmente las observaciones de las variables aleatorias) y es necesario adoptar uno u otro punto de vista

o tomar una decisión determinada sobre la naturaleza del fenómeno sujeto a examen. Ahora bien, aquí se trata de uno de los aspectos más importantes de la actividad humana: el proceso de conocimiento. La tesis de que "el criterio de la verdad es la práctica" está directamente relacionada con la estadística matemática, puesto que precisamente esta ciencia estudia los métodos (en el marco de los modelos matemáticos exactos) que permiten responder a la pregunta de si corresponde o no la práctica, representada en forma de los resultados del experimento, a la referida noción hipotética acerca de la naturaleza del fenómeno.

En este caso es necesario subrayar que, al igual que en la teoría de las probabilidades, nos interesarán no los experimentos que permiten sacar determinadas deducciones unívocas sobre los fenómenos examinados en la naturaleza, sino los experimentos cuyos resultados son sucesos aleatorios. Con el desarrollo de la ciencia, los problemas de tal género desempeñan un papel cada vez más importante, puesto que con el aumento de la precisión de los experimentos es cada vez más difícil evitar el "factor aleatorio" relacionado con diversos tipos de obstáculos y con nuestras limitadas posibilidades de medición y de cálculo.

La estadística matemática forma parte de la teoría de las probabilidades en el sentido de que cada problema de la estadística matemática es, en esencia, un problema (a veces muy peculiar) de la teoría de las probabilidades. Pero la estadística matemática, como tal, también ocupa una posición independiente en la clasificación de las ciencias. La estadística matemática puede considerarse como la ciencia del llamado comportamiento inductivo del hombre (y no sólo del hombre) en condiciones cuando éste, a base de su propia experiencia, debe tomar decisiones con las mínimas pérdidas para él [*].

La estadística matemática también se llama teoría de las decisiones estadísticas, puesto que la misma puede ser caracterizada como la ciencia de las soluciones óptimas (las dos palabras siguientes requieren aclaración) basadas en los datos estadísticos (experimentales). Los planteamientos precisos de los problemas se darán posteriormente en el texto principal del libro. Aquí nos limitaremos a citar tres ejemplos de los problemas estadísticos más elementales y típicos.

Ejemplo 1. Para muchos artículos su plazo de servicio es uno de los parámetros principales que caracteriza la calidad. No obstante, el plazo de servicio de un artículo (digamos, de una bombilla eléctrica) es, por regla general, aleatorio y no se puede determinar de antemano. La experiencia muestra que si el proceso de producción es, en cierto sentido, homogéneo, los plazos de servicio ξ_1, ξ_2 ... de los respectivos artículos 1, 2 etc. pueden

[*] Esta cuestión se examina más detalladamente en [46].

considerarse como magnitudes independientes igualmente distribuidas. El parámetro que nos interesa y que determina el plazo de servicio es natural identificarlo con el número $\theta = M\xi_i$. Uno de los problemas estándar consiste en determinar a qué es igual θ. Para hallar este valor se toman n artículos fabricados y los mismos se someten a comprobación. Sean $x_1, x_2, \ldots, x_n$ los plazos de servicio de dichos artículos comprobados. Sabemos que

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \underset{c.s.}{\to} \theta$$

para $n \to \infty$. Por eso es natural esperar que, al ser n suficientemente grande, el número $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ resultará próximo a θ y permitirá, en cierta medida, responder a las cuestiones planteadas. Es evidente que estamos interesados en que el número requerido de observaciones n sea el menor posible, y que nuestra estimación del número θ sea la más exacta posible (el aumento del parámetro θ, al igual que su reducción, conducirán a pérdidas materiales).

Ejemplo 2. Un radar explora, en los instantes de tiempo $t_1, t_2, \ldots, t_n$, una parte dada del espacio aéreo con el fin de localizar allí cierto objeto. Designemos por $x_1, \ldots, x_n$ los valores de las señales reflejadas que han sido recibidas por el radar. Si en la parte observada del espacio, el objeto que nos interesa no está presente, los valores de x_i pueden considerarse como variables aleatorias independientes distribuidas al igual que cierta variable aleatoria ξ cuya naturaleza está determinada por el carácter de las interferencias diferentes. Pero si en el transcurso de todo el período de observaciones, el objeto se encontraba en el campo de visión, entonces x_i contendrán, al igual que las interferencias, la señal "útil" a, y los valores de x_i se distribuirán como $\xi + a$. Ahora bien, si en el primer caso las observaciones de x_i tenían la función de distribución $F(x)$, en el segundo caso su función de distribución tendrá la forma $F(x - a)$. Por la muestra de $x_1, \ldots, x_n$ es preciso decidir cuál de estos dos casos tiene lugar, o sea, si existe o no, en la parte observada del espacio, el objeto que nos interesa.

En este problema será posible señalar, en cierto sentido, "la regla óptima decisiva" que resolverá el problema planteado, con errores mínimos. No obstante, el problema enunciado puede ser complicado del modo siguiente. Primero falta el objeto y luego, a partir de la observación de número θ desconocido, el mismo aparece. Hay que determinar, lo más exactamente posible, el instante θ de su aparición. Es el llamado "problema de desarreglo" que también tiene una serie completa de otras interpretaciones importantes para su aplicación.

Ejemplo 3. Cierto experimento se realiza al principio n_1 veces en condiciones A y luego n_2 veces en condiciones B. Designemos por $x_1, \ldots, x_{n_1}$ e $y_1, \ldots, y_{n_2}$ los resultados de estos experimentos en condiciones A y B, respectivamente. Es necesario contestar a la pregunta: ¿se reflejará el cambio de las condiciones del experimento en sus resultados? Con otras palabras, si designamos por $\mathbf{P}_A$ la distribución de x_i, $1 \leqslant i \leqslant n_1$, y por $\mathbf{P}_B$, la distribución y_i, $1 \leqslant i \leqslant n_2$, entonces la cuestión consistirá en contestar a la pregunta si se cumplirá o no la relación $\mathbf{P}_A = \mathbf{P}_B$.

Por ejemplo, si hay que determinar si influye o no cierto preparado en el desarrollo, digamos, de las plantas o los animales, entonces paralelamente se hacen dos series de experimentos (con el preparado y sin éste) cuyos resultados es preciso saber compararlos.

A menudo también surgen problemas más complejos cuando una cuestión análoga se plantea para muchas series de observaciones realizadas en condiciones diferentes. Si los resultados de tales observaciones dependen de las condiciones, suele ser necesario comprobar el distinto carácter de esta dependencia (el llamado problema de regresión).

El ejemplo 3 y los problemas más complejos anteriormente mencionados pertenecen a la clase de problemas estadísticos *con dos muestras y más.* Los mismos se examinan en el capítulo 4.

Podríamos continuar la lista de ejemplos de problemas estadísticos típicos, distintos en cuanto a su complejidad y a su esencia. No obstante, para ellos serán comunes las siguientes dos circunstancias:

1. No tendríamos ninguna dificultad si conociéramos las distribuciones de los resultados de las observaciones que figuran en los problemas.

2. En cada uno de estos problemas debemos, a base de los resultados de los experimentos, tomar cierta decisión en cuanto a la distribución de las observaciones disponibles (de aquí precisamente proviene la denominación "Teoría de las resoluciones estadísticas" mencionada más arriba).

En virtud de estas dos advertencias, para la exposición del material ulterior y, en particular, para la resolución de los problemos citados como ejemplos, adquiere importancia de principio el siguiente hecho. Según los resultados de las observaciones $x_1, \ldots, x_n$ de cierta variable aleatoria ξ, es posible, con grandes valores de n, restablecer, tan exactamente como se quiera, la distribución desconocida $\mathbf{P}$ de dicha variable aleatoria. La afirmación análoga también es válida para toda funcional $\theta = \theta(\mathbf{P})$ de esta distribución desconocida.

En este hecho se basa la estadística matemática. A él y a planteamientos más precisos de los problemas está dedicado el capítulo I.

CAPÍTULO I

Muestra. Distribución empírica. Propiedades asintóticas de las estadísticas.

En los §§ 1—4 se introducen los conceptos de muestra y de distribución empírica y se examinan sus propiedades elementales, principalmente asintóticas, que son la base de la estadística matemática.

En el § 5 se exponen los llamados teoremas de continuidad (sobre la convergencia de las distribuciones de las funciones de las sucesiones de variables aleatorias) que se utilizan en todo el libro.

Los §§ 6—10 están dedicados a propiedades asintóticas más finas de las distribuciones empíricas y al estudio de las distribuciones límites para los tipos principales de estadísticas.

§ 1. Concepto de muestra

El conjunto de resultados de las observaciones sirve de material inicial para toda investigación estadística. En los casos elementales, estos resultados no son más que los valores experimentales (obtenidos en las pruebas) de cierta variable aleatoria ξ. Ya hemos señalado que en los problemas de estadística, la distribución $\mathbf{P}$ de esta variable aleatoria se desconoce por lo menos parcialmente.

Supongamos que G es un experimento relacionado con la variable aleatoria ξ. Formalmente, para este experimento debemos construir un modelo matemático del cual forme parte el espacio probabilístico $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mathbf{P})$, y asignarle, de modo conveniente, la función medible que precisamente se denomina variable aleatoria ξ (véase [11]). El espacio $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mathbf{P})$, sin limitar la generalidad, puede considerarse "muestral" (véase [11]), o sea, podemos estimar que $\mathscr{X}$ es el espacio de los valores de $\xi(x) = x$. En este caso $\mathbf{P}$ se puede denominar distribución de ξ.

Si ξ es una variable aleatoria numérica, $\mathscr{X}$ es la recta numérica R; si ξ es un vector, $\mathscr{X} = R^m$, $m > 1$. En lo sucesivo tendremos en cuenta, por regla general, sólo estos dos casos, o sea, por $\mathscr{X}$ entenderemos R (caso uni-

dimensional) o bien R^m, $m > 1$ (caso multidimensional). En calidad de $\mathfrak{B}_{\mathscr{X}}$ se elige con más frecuencia el σ-álgebra de conjuntos de Borel *).

Si se sabe de antemano que $\mathbf{P}$ está concentrada en la parte $B \in \mathfrak{B}_{\mathscr{X}}$ del espacio $\mathscr{X}$, por $\mathscr{X}$ puede resultar cómodo entender B, y por $\mathfrak{B}_{\mathscr{X}}$, la contracción del σ-álgebra $\mathfrak{B}_{\mathscr{X}}$ sobre B.

Examinemos n repeticiones independientes del experimento G (véase [11], p. 38) y designemos por $x_1, \ldots, x_n$ el conjunto de observaciones obtenidas. El vector

$$X_n(x_1, \ldots, x_n)$$

se llama *muestra de volumen n de la población con distribución* $\mathbf{P}$. A veces se utilizan variantes más breves o más completas de este término: "*muestra de la distribución* $\mathbf{P}$" o "*muestra simple de volumen n de la población madre con distribución* $\mathbf{P}$".

Simbólicamente, la relación "$\mathbf{X}_n$ es una muestra de la distribución $\mathbf{P}$" se escribirá, por medio del signo $\in$, del modo siguiente:

$$\mathbf{X}_n \in \mathbf{P}. \tag{1}$$

Tal forma de escritura también será utilizada para otras variables aleatorias. Por ejemplo, la relación

$$\xi \in \mathbf{P} \tag{2}$$

significará que ξ tiene la distribución $\mathbf{P}$. Tal uso del símbolo $\in$ se halla en correspondencia con (1), puesto que esta última ha sido determinada para cualquier n, en particular, para $n = 1$.

Si ξ y η son dos variables aleatorias (dadas, hablando en general, en diferentes espacios) con iguales distribuciones, designaremos este hecho por $\xi \underset{d}{=} \eta$, así que si X_n e Y_n son dos muestras de igual volumen de la distribución $\mathbf{P}$, podemos escribir $X_n \underset{d}{=} Y_n$.

En los segundos miembros de (1) y (2), en vez de la distribución $\mathbf{P}$ puede figurar, a veces, la función de distribución correspondiente a $\mathbf{P}$. Así que si $F(x) = \mathbf{P}((-\infty, x))$, la escritura de

$$X_n \in F$$

será idéntica a (1).

El propio concepto de "muestra de la población madre" también se

*) Muchas partes del libro también serán válidas en una situación más general, cuando $\mathscr{X}$ es un espacio métrico arbitrario con un σ-álgebra $\mathfrak{B}_{\mathscr{X}}$ de conjuntos de Borel, o sea, con un σ-álgebra originada por los conjuntos abiertos de $\mathscr{X}$

encuentra al examinar modelos probabilísticos elementales relacionados con la extracción de bolas de una urna, en la definición clásica de la probabilidad (véase [11], § 2 cap. 1). Cabe señalar que la definición de la muestra, introducida más arriba, se halla en plena correspondencia con este concepto introducido anteriormente y, en esencia, coincide con él. Si x_j (o la variable aleatoria ξ) pueden adoptar sólo s valores $a_1, \ldots, a_s$, y las probabilidades de estos valores son racionales, o sea,

$$\mathbf{P}(\xi = a_j) = \frac{N_j}{N}, \quad \sum_{j=1}^{s} N_j = N,$$

entonces la muestra X_n puede representarse como el resultado del "muestreo con devolución" (en el sentido del cap. 1 [11]) de una urna con N bolas, entre las cuales N_1 bolas están marcadas con a_1, N_2 bolas con a_2, etc.

Como objeto matemático la muestra $X = X_n$ (el índice n será con frecuencia omitido) no es sino la variable aleatoria $(x_1, \ldots, x_n)$ con valores en el espacio "n-dimensional" $\mathscr{X}^n = \mathscr{X} \times \mathscr{X} \times \ldots \times \mathscr{X}$ y con una distribución que para $B = B_1 \times B_2 \times \ldots \times B_n$, $B_j \in \mathfrak{B}_{\mathscr{X}}$ se determina por las igualdades

$$\mathbf{P}(X \in B) = \mathbf{P}(x_1 \in B_1, \ldots, x_n \in B_n) = \prod_{i=1}^{n} \mathbf{P}(x_i \in B_i) \tag{3}$$

Con otras palabras, la distribución $\mathbf{P}$ sobre $\mathscr{X}$ es el producto directo múltiplo de n de las distribuciones "unidimensionales" dadas.

En lo que concierne a las designaciones de la distribución $\mathbf{P}$ y otras, nos sujetaremos a las siguientes acuerdos que ya hemos utilizado parcialmente en (3) y que nunca provocarán equivocaciones.

1. Utilizaremos el mismo símbolo (en particular, $\mathbf{P}$) para las distribuciones en $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ y para el producto directo de estas distribuciones en $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ (véase (3)), donde $\mathfrak{B}^n_{\mathscr{X}}$ es el σ-álgebra de los conjuntos de Borel en $\mathscr{X}^n$. La diferencia será determinada tan sólo por el argumento de la función $\mathbf{P}$.

2. La probabilidad de llegada de la variable X, digamos, de $\mathfrak{B}^n_{\mathscr{X}}$ al conjunto B, a veces será cómodo designarla por $\mathbf{P}(B)$, y a veces por $\mathbf{P}(x \in B)$. Esto es lo mismo, ya que $\mathscr{X}^n$ es el espacio muestral de X.

3. Por último, utilizaremos el símbolo $\mathbf{P}$ para designar el concepto general de probabilidad (o sea, la probabilidad correspondiente a cualesquiera otras variables aleatorias sin concretizar el espacio probabilístico).

En virtud de (3) podemos considerar la muestra X como un suceso elemental en el espacio probabilístico muestral $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}}, \mathbf{P})$ (véase [11] capítulo 3, § 2). Señalemos que en cuanto a la muestra X admitiremos una

interpretación doble de esta designación y del objeto: como variable aleatoria y como vector de los datos numéricos reales obtenidos en los experimentos realmente realizados. Como muestra la experiencia, tal interpretación doble es bien tolerable y no suscita equivocaciones, aunque admite la existencia simultánea de las notaciones que tienen la forma $\mathbf{P}(x_1 < t) = F(t)$ y la forma $x_1 = 0{,}74$, $x_2 = 0{,}83$, etc.

La muestra es el objeto inicial principal en los problemas de la estadística matemática. Sin embargo, en la práctica, sus elementos $x_1, x_2, \ldots$ no siempre, ni mucho menos, son independientes. En nuestros análisis tampoco excluiremos tal posibilidad. Además, para no hacer menciones adicionales, en caso de observaciones dependientes consideraremos que se trata de una muestra de volumen $n = 1$, mientras que las observaciones no son más que las coordenadas del vector x_1 (en efecto, la naturaleza del espacio $\mathscr{X}$ es arbitraria).

En lo sucesivo tendremos que examinar a menudo las muestra X_n de volumen n indefinidamente creciente. En tales casos es cómodo suponer que se da la muestra $X_\infty = (x_1, x_2 \ldots)$ de volumen infinito, y $X = X_n$ no es sino la población de sus primeras n coordenadas. Por muestra de volumen infinito X_∞ entenderemos el elemento del espacio probabilístico muestral $(\mathscr{X}^\infty, \mathfrak{B}_{\mathscr{X}}^\infty, \mathbf{P})$, donde $\mathscr{X}^\infty$ es el espacio de sucesiones $(x_1, x_2, \ldots)$; σ-álgebra $\mathfrak{B}_{\mathscr{X}}^\infty$ ha sido generada por los conjuntos $\bigcap_{i \leqslant N} \{x_i \in B_i\}$, $B_i \in \mathfrak{B}_{\mathscr{X}}$, $N = 1, 2, \ldots$; la distribución $\mathbf{P}$ posee la propiedad (3). Según el teorema de Kolmogórov ([11]), tal distribución siempre existe. Por consiguiente, la suposición sobre la existencia de la muestra X_∞ de volumen infinito de ningún modo limita la generalidad.

La propia sucesión infinita (muestra infinita) X_∞, en los estudios de carácter teórico-probabilístico puede interpretarse como un suceso elemental (compárese con [11]).

En los casos cuando necesitamos entender X_n como un subvector X_∞ escribiremos

$$X_n = [X_\infty]_n,$$

donde $[\cdot]_n$ es el operador de proyección de $\mathscr{X}^\infty$ en $\mathscr{X}^n$, determinado de modo evidente. Con arreglo a lo dicho anteriormente, la notación

$$X_\infty \in \mathbf{P}$$

significará que X_∞ es la muestra de volumen infinito de la distribución $\mathbf{P}$.

Si surge la necesidad de señalar especialmente el hecho de que no se trata de la distribución en $(\mathscr{X}^n, \mathfrak{B}_{\mathscr{X}}^n)$, sino en $(\mathscr{X}^\infty, \mathfrak{B}_{\mathscr{X}}^\infty)$ o en $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ para $n < \infty$, también utilizaremos la designación $\mathbf{P}^\infty$ ($\mathbf{P}^n$). La conservación de los índices superiores "∞" y "n" en todo el texto llevaría a designaciones muy complejas.

§ 2. Distribución empírica (caso unidimensional)

Sea dada la muestra $X = (x_1, \ldots, x_n) \in \mathbf{P}$, $x_i \in \mathscr{X} = R$. Examinemos la recta real R con σ-álgebra de los conjutos de Borel $\mathfrak{B}$ en la distribución discreta $\mathbf{P}_n^*$ sobre $(R, \mathfrak{B})$ concentrada en los puntos $x_1, \ldots, x_n$, para la cual la probabilidad del valor x_i se supone igual a $1/n$. En otros términos, para todo $B \in \mathfrak{B}$, según la definición,

$$\mathbf{P}_n^*(B) = \frac{\nu(B)}{n}, \tag{1}$$

donde $\nu(B)$ es el número de elementos de la muestra X que se encuentran en el conjunto B. La distribución $\mathbf{P}_n^*$ se llama *distribución empírica* construida según la muestra X (o correspondiente a la muestra X). Esta distribución también puede representarse de la forma siguiente. Sea $\mathbf{I}_x(B)$ la distribución concentrada en el punto x:

$$\mathbf{I}_x(B) = \begin{cases} 1, & x \in B, \\ 0, & x \notin B; \end{cases}$$

entonces, evidentemente, $\nu(B) = \sum_{i=1}^{n} I_{x_i}(B)$,

$$\mathbf{P}_n^*(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{x_i}(B). \tag{2}$$

Está claro que para todo B de Borel, $\mathbf{P}_n^*(B)$ como función de la muestra es una variable aleatoria. Ahora bien, se trata de una función aleatoria de los conjuntos, o bien de una distribución aleatoria.

Supongamos ahora que $X_\infty \in \mathbf{P}$, $X_n = [X_\infty]_n$ y $n \to \infty$. Entonces obtendremos una sucesión de distribuciones empíricas $\mathbf{P}_n^*$. El hecho interesante consiste en que esta sucesión se aproxima indefinidamente a la distribución inicial $\mathbf{P}$ de la variable aleatoria sujeta a observación. Este hecho tiene importancia de principio para toda la exposición sucesiva, ya que el mismo muestra que la distribución desconocida $\mathbf{P}$ puede ser restablecida tan exactamente como se quiera, basándose en una muestra de volumen suficientemente grande.

Teorema 1. *Sea* $B \in \mathfrak{B}$ *y* $X_n = [X_\infty]_n \in \mathbf{P}$. *Entonces, para* $n \to \infty$

$$\mathbf{P}_n^*(B) \underset{\text{c.s.}}{\to} \mathbf{P}(B).$$

La convergencia con la probabilidad 1 aquí se sobreentiende con respecto a la distribución $\mathbf{P} = \mathbf{P}^\infty$ en $(R^\infty, \mathfrak{B}^\infty, \mathbf{P})$. Necesitamos la suposición $X_n = [X_\infty]_n$ para que las variables aleatorias $\mathbf{P}_n^*(B)$ se den en un solo espacio probabilístico.

Demostración. Examinemos la definición (2) y notemos que $\mathbf{I}_{x_i}(B)$ son variables aleatorias independientes igualmente distribuidas, $\mathbf{MI}_{x_i}(B) = \mathbf{P}(\mathbf{I}_{x_i}(B) = 1) = \mathbf{P}(x_i \in B) = \mathbf{P}(B)$. Como $\mathbf{P}_n^*(B)$ es la media aritmética de estas variables, nos queda hacer uso de la ley fuerte de los grandes números. ◁

El teorema 1 establece la convergencia de $\mathbf{P}_n^*(B)$ y $\mathbf{P}(B)$ en cada "punto" de B. No obstante, también tiene lugar una afirmación más fuerte de que tal convergencia es, en cierto sentido, uniforme respecto a B.

Designemos por $\mathfrak{F}$ la población de los conjuntos B que son semiintervalos de forma $[a, b)$ con extremos finitos o infinitos y volvamos a suponer que $X_n = [X_\infty]_n$.

Teorema 2 (de Glivenko — Cantelli).

$$\sup_{B \in \mathfrak{F}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{c.s}}{\to} 0.$$

A decir verdad, con los nombres de Glivenko y Cantelli está relacionada una afirmación algo diferente, que se refiere a un concepto importante de la *función empírica de distribución*. Por definición, ésta es la función de distribución correspondiente a $\mathbf{P}_n^*$. En otros términos, se llama *función empírica de distribución* $F_n^*(x)$ la función

$$F_n^*(x) = \mathbf{P}_n^*((-\infty, x)).$$

La variable $nF_n^*(x)$ es igual al número de elementos de la muestra que son menores que x. En las condiciones reales, para construir $F_n^*(x)$ se utiliza a menudo el procedimiento siguiente. Los elementos de la muestra $(x_1, \ldots, x_n)$ se ordenan de manera creciente, o sea, de ella se forma la sucesión

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

que se llama *serie variacional*. Entonces podemos suponer que

$$F_n^*(x) = \frac{k}{n} \text{ para } x \in (x_{(k)}, x_{(k+1)}),$$

donde k recorre los valores de 0 a n, $x_{(0)} = -\infty$, $x_{(n+1)} = \infty$. Evidentemente, $F_n^*(x)$ es una función escalonada que tiene saltos de $1/n$ en los puntos x_i si todos los valores de x_i son diferentes.

Sea $F(x) = \mathbf{P}(-\infty, x)$ la función de la distribución ξ (o x_1, que es lo mismo) y $X_n = [X_\infty]_n$. El teorema de Glivenko — Cantelli consiste en lo siguiente:

Teorema 2A. *Si* $n \to \infty$

$$\sup_x |F_n^*(x) - F(x)| \underset{\text{c.s.}}{\to} 0.$$

Más abajo omitiremos el índice n en las designaciones de F_n^* y escribiremos simplemente F^*.

Demostración del teorema 2A. Para abreviar supongamos primeramente que la función F es continua. Sea $\varepsilon > 0$ un número dado, arbitrariamente pequeño, de tal modo que el número $N = 1/\varepsilon$ sea entero. Cómo F es continua, podemos señalar los números $z_0 = -\infty, z_1, \ldots, z_{N-1}, z_N = \infty$ con los que

$$F(z_0) = 0,\ F(z_1) = \varepsilon = \frac{1}{N},\ \ldots,\ F(z_k) = k\varepsilon = \frac{k}{N},\ \ldots$$
$$\ldots,\ F(z_N) = 1.$$

Para $z \in [z_k,\ z_{k+1})$ son válidas las relaciones

$$\begin{aligned} F^*(z) - F(z) &\leqslant F^*(z_{k+1}) - F(z_k) = F^*(z_{k+1}) - F(z_{k+1}) + \varepsilon, \\ F^*(z) - F(z) &\geqslant F^*(z_k) - F(z_{k+1}) = F^*(z_k) - F(z_k) - \varepsilon. \end{aligned} \tag{3}$$

Designemos por A_k el conjunto de sucesos elementales $\omega = X_\infty$ en los cuales $F^*(z_k) \to F(z_k)$. Según el teorema 1, $\mathbf{P}(A_k) = 1$. Por consiguiente, para cada $\omega \in A = \bigcap_{k=0}^{N} A_k$ se encontrará un valor de $n(\omega)$ tal, que para todos los valores de $n \geqslant n(\omega)$ se cumplirá

$$|F^*(z_k) - F(z_k)| < \varepsilon, \quad k = 0,\ 1,\ \ldots,\ N. \tag{4}$$

Pero junto con (3), dichas desigualdades contribuyen a que

$$\sup_z |F^*(z) - F(z)| < 2\varepsilon. \tag{5}$$

Así pues, esta relación tiene lugar para un valor arbitrario de $\varepsilon > 0$, para todos los valores de $\omega \in A$ y para todos los valores bastante grandes de $n \geqslant n(\omega)$. Como $\mathbf{P}(A) = 1$, el teorema para la función continua F se considera demostrado.

Para la función arbitraria $F(x)$, la demostración del teorema se realiza absolutamente igual. Se debe sólo hacer uso de la circunstancia siguiente: para toda $F(x)$ existe un número finito de puntos $-\infty = z_0 < z_1 < \ldots$ $\ldots < z_{N-1} < z_N = \infty$ con los que

$$F(z_{k+1}) - F(z_k + 0) \leqslant \varepsilon, \quad k = 0,\ 1,\ \ldots,\ N - 1 \tag{6}$$

(para evidenciar podemos considerar que el conjunto $\{z_j\}$ contiene todos los puntos de los saltos de F que por sus valores superan, por ejemplo, $\varepsilon/2$). Absolutamente igual que en (3) obtenemos que para $z \in (z_k,\ z_{k+1}]$,

$$\begin{aligned} F^*(z) - F(z) &\leqslant F^*(z_{k+1}) - F(z_{k+1}) + \varepsilon, \\ F^*(z) - F(z) &\geqslant F^*(z_k + 0) - F(z_k + 0) - \varepsilon. \end{aligned} \tag{7}$$

A los conjuntos A_k, que se determinan como antes, les agregaremos los conjuntos A_k^+, $k = 0, 1, \ldots, N$ en los que $F^*(z_k + 0) \to F(z_k + 0)$. Entonces, según el teorema 1, $\mathbf{P}(A_k) = \mathbf{P}(A_k^+) = 1$, y en el conjunto $A = \bigcap_{k=0}^{N} A_k A_k^+$, $\mathbf{P}(A) = 1$, para valores de $n \geqslant n(\omega)$ bastante grandes será válida (4), así como las desigualdades

$$|F^*(z_k + 0) - F(z_k + 0)| < \varepsilon, \quad k = 0, 1, \ldots, N.$$

Junto con (7) estas desigualdades conducen a (5). ◁

El teorema 2A es un caso particular del teorema 2, ya que los conjuntos $(-\infty, x)$ pertenecen a $\mathfrak{F}$; por otro lado, el teorema 2 se obtiene fácilmente en calidad de corolario del teorema 2A, puesto que para $B = [a, b)$

$$|\mathbf{P}_n^*(B) - \mathbf{P}(B)| \leqslant |F_n^*(b) - F(b)| + |F_n^*(a) - F(a)|,$$

y, por consiguiente,

$$\sup_{B \in \mathfrak{F}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \leqslant \sup_{a,\, b} [|F_n^*(b) - F(b)| + |F_n^*(a) - F(a)|] \underset{\text{c.s.}}{\to} 0.$$

Observación 1. Es fácil notar que los razonamientos de ese mismo género nos permiten, en calidad de población de los conjuntos $\mathfrak{F}$ en el teorema 2, tomar las poblaciones de los intervalos (a, b), de los segmentos $[a, b]$ y de sus uniones finitas (de número no mayor que cierto N).

Por otro lado, si en calidad de $\mathfrak{F}$ en el teorema 2 se toma una clase bastante rica de conjuntos, la afirmación del teorema deja de ser justa. Por ejemplo, si $\mathfrak{F}$ contiene las uniones de cualquier número finito de intervalos, entonces el conjunto $B_n = \bigcup_{k=1}^{n} (x_k - 1/n^2, \; x_k + 1/n^2) \in \mathfrak{F}$, $\mathbf{P}_n^*(B_n) = 1$ y para la distribución uniforme en $[0, 1]$, $\mathbf{P}(B_n) \leqslant 2/n$, así que

$$\sup_{B \in \mathfrak{F}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \geqslant \mathbf{P}_n^*(B_n) - \mathbf{P}(B_n) \to 1.$$

Concluyendo este párrafo señalaremos que la representación (2) permite obtener para $\mathbf{P}_n^*$ teoremas sobre el comportamiento asintótico aún más exactos que los teoremas del tipo de Glivenko — Cantelli (estos resultados serán representados en los §§ 4 y 6). Para ilustrar las posibilidades que aquí existen recordemos que $\sum_{i=1}^{n} \mathbf{I}_{x_i}(B)$ en (2) es la suma de las variables aleatorias independientes e igualmente distribuidas en el esquema de Bernoulli

$$\mathbf{M}\mathbf{I}_{x_i}(B) = \mathbf{P}(\mathbf{I}_{x_i}(B) = 1) = \mathbf{P}(B),$$
$$\mathbf{M}\mathbf{I}_{x_i}^2(B) = \mathbf{P}(B), \quad \mathbf{D}\mathbf{I}_{x_i}(B) = \mathbf{P}(B)(1 - \mathbf{P}(B)).$$

Por eso, del teorema central del límite se deduce inmediatamente la afirmación siguiente:

Teorema 3. $\mathbf{P}_n^*(B)$ *es representable en la forma*

$$\mathbf{P}_n^*(B) = \mathbf{P}(B) + \frac{\zeta_n(B)}{\sqrt{n}}, \tag{8}$$

donde la distribución $\zeta_n(B) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{I}_{x_i}(B) - \mathbf{P}(B))$ *converge hacia la distribución normal con los parámetros* $(0, \mathbf{P}(B)(1 - \mathbf{P}(B))$.

El estudio ulterior de $\mathbf{P}_n^*(B)$ en este sentido se ofrece en el § 6. Teoremas más exactos sobre la convergencia con probabilidad 1 se dan en el § 4.

§ 3. Características muestrales. Dos tipos de estadísticas

1. Ejemplos de características muestrales. Por *características muestrales* suelen entenderse las diversas funcionales medibles de una distribución empírica o, dicho de otro modo, las funciones de una muestra que se supone que son medibles. Entre ellas, los momentos muestrales (o empíricos) son los más simples. Llámase *momento muestral de orden k* el valor de

$$a_k^* = a_k^*(X) = \int x^k dF_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

El momento central muestral de orden k es igual a

$$a_k^{*\circ} = a_k^{*\circ}(X) = \int (x - a_1^*)^k dF_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - a_1^*)^k.$$

Para los momentos muestrales a_1^* y $a_2^{*\circ}$, en la literatura se utilizan designaciones especiales, $\bar{x}$ y S^2:

$$\bar{x} = a_1^* = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad S^2 = a_2^{*\circ} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

En los problemas estadísticos se usan las características muestrales más diferentes. Por ejemplo, la *mediana muestral* ζ^* es el valor medio de una serie variacional, o sea, el valor de $\zeta^* = x_{(m)}$ si $n = 2m - 1$ (impar) y $\zeta^* = (x_{(m)} + x_{(m+1)})/2$ si $n = 2m$ (par). Recordemos que por mediana ζ de la distribución continua $\mathbf{P}$ se entiende la solución de la ecuación $F(\zeta) = 1/2$.

Un concepto más general es el de *cuantila* ζ_p de orden p. Es el número para el cual $F(\zeta_p) = p$. Así que la mediana es una cuantila de orden 1/2. Si F tiene puntos de discontinuidad (componente discreta) entonces esta definición pierde su sentido. Por eso en un caso general utilizaremos la definición siguiente:

Se denomina *cuantila* ζ_p *de orden* p de la distribución $\mathbf{P}$ el número

$$\zeta_p = \sup \{x: F(x) \leqslant p\}.$$

Como función de p la cuantila ζ_p no es más que la función $F^{-1}(p)$, inversa a $F(x)$.

Es evidente que, a diferencia de la anterior, esta definición de ζ_p (o de $F^{-1}(p)$) tiene sentido para cualesquiera $F(x)$.

Es natural que a la par con las medianas muestrales podemos examinar las *cuantilas muestrales* ζ_p^* *de orden* p que por definición son iguales al valor de $x_{(l)}$, donde $l = [np] + 1$, $x_{(k)}$ son los términos de la serie variacional para la muestra X, $k = 1, \ldots, n$. Para $p = 1/2$ utilizaremos la definición $\zeta^* = \zeta_{1/2}^*$ que hemos dado anteriormente (coincide tan sólo con la definición dada para n impares).

2. Dos tipos de estadísticas. Supongamos que se da una función medible S de n argumentos. La característica muestral $S(X) = S(x_1, \ldots, x_n)$ a menudo también se llama *estadística*. De lo dicho anteriormente se deduce que cualquier estadística es una variable aleatoria. Su distribución se determina por completo mediante la distribución $\mathbf{P}(B) = \mathbf{P}(x_1 \in B)$ (recordemos que $S(X)$ se puede considerar como una variable aleatoria dada en $(\mathscr{X}^n, \mathfrak{B}_{\mathscr{X}}^n, \mathbf{P})$, donde $\mathbf{P}$ es el producto directo múltiplo de n de las distribuciones unidimensionales de x_1).

Destaquemos aquí dos clases de características que se encontrarán frecuentemente a continuación. Se construirán con ayuda de los dos tipos siguientes de funcionales $G(F)$ de las funciones de distribución F:

I. Funcionales que tienen la forma

$$G(F) = h\left(\int g(x)dF(x)\right),$$

donde g es la función dada de Borel; h, la función continua en el punto $a = \int g(x)dF_0(x)$, donde F_0 es tal que $X \in F_0$.

II. Funcionales $G(F)$ continuas en el "punto" F_0 en la métrica uniforme: $G(F^{(n)}) \to G(F_0)$, si $\sup_x |F^{(n)}(x) - F_0(x)| \to 0$, los portadores *) de las distribuciones de $F^{(n)}$ pertenecen al portador de F_0. Aquí, como antes, F_0 es la función para la cual $X \in F_0$.

*) El portador N_F de la distribución $\mathbf{P}$ con la función de distribución F es el conjunto para el cual $\mathbf{P}(N_F) = 1$.

Vamos a definir las clases respectivas de estadísticas con ayuda de la igualdad

$$S(X) = G(F_n^*),$$

donde F_n^* es la función empírica de distribución. Entonces obtenemos:

I. Clase de *estadísticas de tipo* I, representables en la forma

$$S(X) = h\left(\int g(x)dF_n^*(x)\right) = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right).$$

Es evidente que todos los momentos muestrales tienen la forma de las estadísticas aditivas $\frac{1}{n}\sum_{i=1}^{n} g(x_i)$ y figuran entre las estadísticas del tipo I.

II. Clase de estadísticas que llamaremos *estadísticas de tipo* II o bien estadísticas *continuas en el punto* F_0.

Está claro que, por ejemplo, la mediana muestral será la estadística continua en el punto F si existe la mediana ζ, $F(\zeta) = 1/2$ y F es continua y crece estrictamente en el punto ζ.

La pertenencia de las funcionales a una de las clases mencionadas no es, desde luego, alternativa. La funcional $G(F)$ puede no pertenecer a ninguna de estas clases o pertenecer a ambas clases a la vez. Por ejemplo, si G es una funcional de tipo I, el portador de F está concentrado en el segmento $[a, b]$ $(F(a) = 0, F(b) = 1)$ y la función g tiene una variación limitada en $[a, b]$, entonces G será simultáneamente una funcional de tipo II, ya que en este caso la funcional

$$\int g(x)dF(x) = g(b) - \int_a^b F(x)dg(x)$$

es continua con respecto a F en la métrica uniforme. Lo dicho quiere decir que las estadísticas de tipo I $\bar{x}$ y S^2 serán también de tipo II si $X \in \mathbf{P}$ y $\mathbf{P}$ está concentrada en el intervalo finito.

Podemos completar los teoremas 2.1 y 2.2 con la siguiente afirmación sobre la convergencia casi segura de las características muestrales.

Teorema 1. *Sea, como antes,* $X_n = |X_\infty|_n \in F$. *En este caso, si* $S(X) = G(F_n^*)$ *es la estadística de tipo* I ó II, *para* $n \to \infty$

$$G(F_n^*) \underset{\text{c.s.}}{\longrightarrow} G(F).$$

Aquí se supone, desde luego, que el valor de $G(F)$ existe.

Ahora bien, las muestras de gran volumen permiten estimar no sólo la propia distribución $\mathbf{P}$, sino también las funcionales de esta distribución,

por lo menos aquellas que pertenecen a una de las clases citadas en el teorema.

Demostración de la afirmación para ambas clases de estadísticas es casi evidente. Sea, por ejemplo, $G(F) = h\left(\int g(x)dF(x)\right)$. Entonces

$$S = S(X) = \int g(x)dF_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

es la suma de las variables aleatorias independientes, con la esperanza matemática

$$\mathbf{M}g(x_1) = \int g(x)dF(x).$$

Por eso en consonancia con la ley fuerte de los grandes números $S \xrightarrow[\text{c.s.}]{} \mathbf{M}g(\mathrm{x}_1)$. Sea ahora $A = \{X_\infty: S(X) \to \mathbf{M}g(\mathrm{x}_1)\}$. Entonces $\mathbf{P}(A) = 1$ y si $X_\infty \in A$, entonces $S(X) \to \mathbf{M}g(\mathrm{x}_1)$, $h(S(X)) \to h(\mathbf{M}g(\mathrm{x}_1))$. Con otras palabras, en el conjunto A

$$G(F_n^*) \to G(F).$$

La afirmación del teorema para las funcionales de segundo tipo es el corolario directo del teorema de Glivenko — Cantelli. ◁

Del teorema se deduce que los momentos absolutos y centrales convergen casi seguramente para $n \to \infty$ a los momentos correspondientes de la distribución $\mathbf{P}$:

$$a_k^* = a_k^*(X) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i^k \xrightarrow[\text{c.s}]{} \mathbf{M}\mathrm{x}_1^k,$$

$$a_k^{*\circ} = a_k^{*\circ}(X) = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{x}_i - \bar{\mathrm{x}})^k \xrightarrow[\text{c.s}]{} \mathbf{M}(\mathrm{x}_1 - \mathbf{M}\mathrm{x}_1)^k.$$

En particular,

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathrm{x}_i - \bar{\mathrm{x}})^2 = \frac{1}{n}\sum \mathrm{x}_i^2 - \bar{\mathrm{x}}^2 \to \mathbf{D}\mathrm{x}_1.$$

Ahora bien, hemos establecido un hecho importante que tiene para nosotros el valor de principio: con el aumento del volumen de la muestra, la distribución empírica y una amplia clase de funcionales de ésta se aproximan indefinidamente a los valores "teóricos" correspondientes.

Teoremas más exactos de la distribución de las características muestrales se exponen en los §§ 7 y 8.

§ 4. Muestras multidimensionales

1. Distribuciones empíricas. De un modo completamente análogo se construyen las distribuciones empíricas y las características muestrales en el caso multidimensional cuando la variable aleatoria observada ξ, y junto con ella también los valores muestrales $x_1, \ldots, x_n$, son vectores de dimensión $m > 1$: $x_k = (x_{k,1}, \ldots, x_{k,m})$. Aquí $\mathbf{P}(B) = \mathbf{P}(\xi \in B)$ es la distribución en $\mathscr{X} = R^m$, y el espacio muestral aquí será $(\mathscr{X}^n, \mathfrak{B}_{\mathscr{X}}^n, \mathbf{P})$, donde $\mathbf{P}$ es el producto directo múltiplo de n de las distribuciones $\mathbf{P}$ en $(R^m, \mathfrak{B}_{\mathscr{X}} = \mathfrak{B}_R^m)$. La designación $X \in \mathbf{P}$ conserva por completo su sentido.

La distribución empírica $\mathbf{P}_n^*$, basada en la muestra X, se construye, al igual que antes, como una distribución discreta con masas de valores $1/n$ en los puntos $x_1, \ldots, x_n$, así que

$$\mathbf{P}_n^*(B) = \frac{\nu(B)}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{x_i}(B),$$

donde $\nu(B)$ es el número de puntos que entran en el conjunto B, y $\mathbf{I}_{x_i}$ la distribución concentrada en el punto x_i.

Es evidente que la afirmación del teorema 1 acerca de la convergencia de $\mathbf{P}_n^*(B) \underset{\text{c.s.}}{\longrightarrow} \mathbf{P}(B)$ aquí también será válida.

La generalización del teorema de Glivenko — Cantelli para el caso multidimensional está relacionada con la aparición de cuestiones cualitativamente nuevas. Una de ellas consiste en generalizar el concepto de intervalos para el caso multidimensional. Puede haber varias generalizaciones de tal género, por ejemplo, rectángulos, conjuntos convexos, etc.

Una variante elemental de generalización del teorema de Glivenko — Cantelli es la siguiente.

Sea $y = (y_1, \ldots, y_m)$ el punto R^m, y B_t, un ángulo con vértice en el punto $t = (t_1, \ldots, t_m)$:

$$B_t = \{y \in R^m: \quad y_k < t_k, \quad k = 1, \ldots, m\}.$$

La función

$$F_n^*(t) = \mathbf{P}_n^*(B_t)$$

se llama *función empírica de distribución.*

Teorema 1. *Sea* $X_n = [X_\infty]_n$, $X_\infty \in F$. *Entonces*

$$\sup_t |F_n^*(t) - F(t)| \underset{\text{c.s.}}{\longrightarrow} 0$$

si $n \to \infty$.

2°. Variantes más generales del teorema de Glivenko — Cantelli. Ley de logaritmo repetido. Una de las generalizaciones posibles de los teoremas

del tipo de Glivenko — Cantelli consiste en lo siguiente. Sea $\mathfrak{C}$ la clase de todos los conjuntos convexos sobre R^m.

Teorema 2. *Supongamos que* $X_n = [X_\infty]_n$, $X_\infty \in \mathbf{P}$ *y que la distribución* $\mathbf{P}$ *es absolutamente continua respecto a la medida de Lebesgue en* R^m. *Entonces*

$$\sup_{B \in \mathfrak{C}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{c.s.}}{\to} 0. \tag{1}$$

Otras generalizaciones posibles del teorema 1 pueden ser obtenidas con ayuda de las afirmaciones del Suplemento 1.

Observación 1. La exigencia de que la distribución $\mathbf{P}$ sea absolutamente continua con respecto a la medida de Lebesgue es muy importante en el teorema 2. Esto lo demuestra el ejemplo siguiente. Sea $\mathbf{P}$ la distribución uniforme en una circunferencia unitaria (o sea, en el límite de un círculo) en R^2. Construyamos el polígono cerrado B_X con los vértices en los puntos $x_1, \ldots, x_n$ situados en dicha circunferencia. Es un conjunto convexo. Sin embargo, $\mathbf{P}(B_X) = 0$, $\mathbf{P}_n^*(B_X) = 1$, es incorrecta y, por consiguiente, también lo es la relación (1), donde $\mathfrak{C}$ es la clase de los conjuntos convexos.

Las afirmaciones de los teoremas del tipo de Glivenko — Cantelli pueden ser precisadas considerablemente, por lo menos, para las clases elementales de conjuntos. Por ejemplo, para las funciones empíricas de distribuciones $F_n^*(t)$ (véase el teorema 1) se puede señalar la siguiente sucesión determinada: $b_n \to 0$ cuando $n \to \infty$, para la cual, con la probabilidad 1 (para casi todos los "puntos" X_∞),

$$\lim_{n\to\infty} \sup b_n^{-1} \sup_t |F_n^*(t) - F(t)| = 1.$$

Resulta que el orden de pequeñez de b_n equivale al de $\sqrt{\dfrac{\ln \ln n}{n}}$.

Teorema 3 (ley del logaritmo repetido). *Si* $F(t)$ *es continua, entonces*

$$\mathbf{P}\left(\lim_{n\to\infty} \sup \sqrt{\frac{2n}{\ln \ln n}} \sup_t |F_n^*(t) - F(t)| = 1\right) = 1.$$

El teorema 3 está estrechamente relacionado con la aproximación normal para $F_n^*(t)$ de la forma (2.8) que, evidentemente, en el caso multidimensional también tiene lugar.

La demostración de los teoremas 1 y 2 se da en el Suplemento 1, y la demostración del teorema 3 véase en [52].

3. Características muestrales. En el caso multidimensional, al igual que en el unidimensional, éstas son distintas funciones medibles de la muestra. Las más elementales de ellas son los momentos muestrales. Por ejemplo, los momentos muestrales de primer orden son iguales a

$$a_{1,j}^{*} = a_{1,j}^{*}(X) = \frac{1}{n}\sum_{k=1}^{n} x_{k,j}, \quad j = 1, \ldots, m.$$

Los momentos de segundo orden (ordinarios y centrales)

$$a_{2,ij}^{*} = a_{2,ij}^{*}(X) = \frac{1}{n}\sum_{k=1}^{n} x_{k,i}x_{k,j}; \quad i, j = 1, \ldots, m,$$

$$a_{2,ij}^{*0} = S_{ij} = \frac{1}{n}\sum_{k=1}^{n} (x_{k,i} - a_{1,i}^{*})(x_{k,j} - a_{1,j}^{*}),$$

etc. Al igual que en el caso unidimensional, con ayuda de la ley fuerte de los grandes números es fácil cerciorarse de que estas características convergen, con probabilidad 1, hacia los momentos "teóricos" correspondientes. En particular, $S_{ij} \underset{\text{c.s.}}{\to} \mathbf{M}(x_{1,i} - \mathbf{M}x_{1,i})(x_{1,j} - \mathbf{M}x_{1,j})$. Es fácil convencerse (esto se analiza más detalladamente en el párrafo siguiente) de que los coeficientes de correlación muestrales

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \underset{\text{c.s.}}{\to} \varrho(x_{1,i}x_{1,j}) = \frac{\mathbf{M}(x_{1,i} - \mathbf{M}x_{1,i})(x_{1,j} - \mathbf{M}x_{1,j})}{\sqrt{\mathbf{D}x_{1,i}\mathbf{D}x_{1,j}}}$$

también poseen esta misma propiedad.

Para obtener teoremas más exactos de la distribución de las características muestrales nos serán útiles los llamados teoremas de continuidad.

§ 5. Teoremas de continuidad

En lo sucesivo necesitaremos ciertos conceptos auxiliares que utilizaremos a menudo y que podrían ser llamados *teoremas de continuidad*. Para facilitar su estudio, a ellos les dedicamos un párrafo especial. Anteriormente ya hemos utilizado un teorema de este tipo — el teorema 3.1. El primer teorema de continuidad será muy parecido a éste.

Teorema 1 (primer teorema de continuidad). *Sea* $X = |X_\infty|_n \in \mathbf{P}$. *En este caso, si* $S_n = S_n(X)$ *es una sucesión de estadísticas escalares o vectoriales, tales que* $S_n \underset{\text{c.s.}}{\to} S_0$, *y* $H(s)$ *es una función continua casi por doquier con respecto a la distribución de la variable aleatoria* S_0 (*o sea,* $H(s)$ *es continua en cada punto del conjunto* B, $\mathbf{P}(S_0 \in B) = 1$), *entonces* $H(S_n(X)) \underset{\text{c.s.}}{\to} H(S_0)$.

Si S_n *converge hacia* S_0 *según la probabilidad* ($S_n \underset{P}{\to} S_0$), *entonces para las demás condiciones semejantes,* $H(S_n) \underset{P}{\to} H(S_0)$.

La demostración del teorema es casi evidente. Como las probabilidades de los sucesos $A = \{X_\infty\colon S_n(X_\infty) \to S_0(X_\infty)\}$ y $C = \{X_\infty\colon S_0(X_\infty) \in B\}$ son iguales a 1, entonces, en virtud de la igualdad $\mathbf{P}(A \cap C) = \mathbf{P}(A) + \mathbf{P}(C) - \mathbf{P}(A \cup C)$, la probabilidad del suceso $A \cap C$ (en el cual $H(S_n(X_\infty)) \to H(S_0(X_\infty))$) también es igual a 1.

Para simplificar la demostración de la convergencia en probabilidad, supongamos adicionalmente que $S_0 = \text{const}$ (sólo necesitaremos este caso). Para un valor dado de $\varepsilon > 0$ hay un valor de $\delta > 0$ tal, que el suceso $A_n = \{X_\infty\colon |S_n - S_0| < \delta\}$ contribuye a que $|H(S_n) - H(S_0)| < \varepsilon$ y además, $\mathbf{P}(A_n) > 1 - \varepsilon$ para todos los valores de n bastante grandes. Por lo tanto, para tales n tenemos $1 - \varepsilon < \mathbf{P}(A_n) \leqslant \mathbf{P}(|H(S_n) - H(S_0)| < \varepsilon)$. ◁

Antes de enunciar los teoremas siguientes, introduzcamas ciertas designaciones que serán cómodas posteriormente.

Supongamos que se ha dado una sucesión de vectores aleatorios $\eta_n = (\eta_n^{(1)}, \ldots, \eta_n^{(s)})$ (no obligatoriamente en el mismo espacio probabilístico). Si las distribuciones η_n convergen débilmente (cuando $n \to \infty$) hacia la distribución de cierta variable aleatoria η, entonces designaremos este hecho con el símbolo

$$\eta_n \Rightarrow \eta. \tag{1}$$

Aquí utilizamos, para las variables aleatorias, el signo $\Rightarrow$ de convergencia débil de las distribuciones. Al igual que antes, utilizaremos también este signo para las propias distribuciones, así que la relación (1) es equivalente a que

$$\mathbf{Q}_n \Rightarrow \mathbf{Q},$$

donde $\mathbf{Q}_n$ y $\mathbf{Q}$ son las distribuciones de η_n y η respectivamente. Tal convenio es cómodo y no conduce a equivocaciones.

Está claro que de $\eta_n \xrightarrow[P]{} \eta$ o de $\eta_n \xrightarrow[\text{c.s.}]{} \eta$ se deduce $\eta_n \Rightarrow \eta$ (compárese con [11], p. 133).

Ahora bien, si se trata de la relación (correspondiente a una convergencia débil) entre objetos de igual naturaleza (entre variables aleatorias o entre distribuciones), usaremos el símbolo $\Rightarrow$. También sería conveniente tener el símbolo para expresar el hecho de que "*las distribuciones de η_n convergen débilmente hacia* $\mathbf{Q}$ *cuando* $n \to \infty$". Escribiremos esta relación de la forma

$$\eta_n \Subset \mathbf{Q}, \tag{2}$$

así que el símbolo $\Subset$ expresa el mismo hecho que $\Rightarrow$, pero une objetos de distinta naturaleza, al igual que el símbolo $\in$ respecto a $\eta \Subset \mathbf{Q}$ (a la izquierda en (2) se encuentran las variables aleatorias, y a la derecha, la distribución).

Sean η_n y η vectores aleatorios de R^s.

Teorema 2 (segundo teorema de continuidad). *Si* $\eta_n \Rightarrow \eta$ *y* $H(t)$, $t \in R^s$ *es una función continua de* R^s *en* R^k, *entonces* $H(\eta_n) \Rightarrow H(\eta)$.

Señalemos que, en realidad, este teorema también es cierto en una forma más general [*]. *Si* $\eta_n \Rightarrow \eta$ *y* $H(t)$ *es continua en los puntos del conjunto* $A \in \mathfrak{B}^s$, $\mathbf{P}(\eta \in A) = 1$, *entonces* $H(\eta_n) \Rightarrow H(\eta)$.

Demostración del teorema 2. Sean $\mathbf{Q}_n$ y $\mathbf{Q}$ las distribuciones η_n y η, respectivamente. La convergencia débil de $\mathbf{Q}_n \Rightarrow \mathbf{Q}$ significa, por definición, que para toda función continua y limitada f: $R^s \to R$ se cumple

$$\int f(y)\mathbf{Q}_n(dy) \to \int f(y)\mathbf{Q}(dy)$$

o bien, que es lo mismo,

$$\mathbf{M}f(\eta_n) \to \mathbf{M}f(\eta). \tag{3}$$

También debemos obtener una relación análoga para las distribuciones $H(\eta_n)$ y $H(\eta)$. O sea, debemos establecer que para toda función continua limitada g: $R^k \to R$ es válida $\mathbf{M}g(H(\eta_n)) \to \mathbf{M}g(H(\eta))$. Pero esto se deduce con evidencia de (3), ya que la superposición $\tilde{g} \equiv g \circ H$: $R^s \to R$ es continua y limitada. ◁

Teorema 3 (tercer teorema de continuidad). *Sea* $\eta_n \Rightarrow \eta \in R$, $H(t)$, $t \in R$ *una función derivable en el punto* a. *Entonces, si* $b_n \to 0$ *es una sucesión numérica,*

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta H'(a). \tag{4}$$

Demostración. Examinemos la función

$$h(x) = \begin{cases} (H(a + x) - H(a))/x, & x \neq 0, \\ H'(a), & x = 0, \end{cases}$$

la cual será continua en el punto $x = 0$. Como $b_n\eta_n \Rightarrow 0$, en virtud del primer teorema de continuidad, $h(b_n\eta_n) \Rightarrow h(0) = H'(a)$. Utilizando el segundo teorema de continuidad, obtenemos

$$(H(a + b_n\eta_n) - H(a))/b_n = h(b_n\eta_n)\eta_n \Rightarrow H'(a)\eta. \quad \triangleleft$$

Ahora citaremos dos generalizaciones sucesivas del teorema 3 para el caso multidimensional, las cuales nos serán útiles.

Teorema 3A. *Supongamos que* $\eta_n \equiv (\eta_n^{(1)}, \ldots, \eta_n^{(s)}) \Rightarrow \eta \equiv (\eta^{(1)}, \ldots, \eta^{(s)})$ *y que* $H(t)$ *es función escalar del vector* $t = (t_1, \ldots, t_s)$ *con la que existe la derivada* $H'(t) \equiv \left(\dfrac{\partial H}{\partial t_1}, \ldots, \dfrac{\partial H}{\partial t_s}\right)$ *en el punto* a. *Entonces, cuando* $b_n \to 0$,

[*] Véase [5].

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta(H'(a))^T = \sum_{j=1}^{s} \frac{\partial H(a)}{\partial t_j}\,\eta^{(j)}. \qquad (5)$$

Aquí el índice T corresponde a la transposición.

Si $\eta(H'(a))^T = 0$ *con probabilidad* 1 *(por ejemplo,* $H'(a) = 0$*), y la matriz* $H''(t)$ *de las derivadas* $\frac{\partial^2 H(t)}{\partial t_i \partial t_j}$ *existe en el punto a, entonces*

$$(H(a + b_n\eta_n) - H(a))/b_n^2 \Rightarrow \frac{1}{2}\,\eta H''(a)\eta^T = \frac{1}{2}\sum_{i,j=1}^{s} \frac{\partial^2 H(a)}{\partial t_i \partial t_j}\,\eta^{(i)}\eta^{(j)}. \qquad (6)$$

Sea ahora $H(t)$ una función vectorial. Entonces, evidentemente, la distribución límite para cada componente H_j será descrita por el teorema 3A, y con respecto a la distribución conjunta será valida.

Teorema 3B. *Supongamos que* $\eta_n \Rightarrow \eta \in R^s$ *y que* $H(t) \in R^k$ *es una función vectorial con la que las derivadas* H_j', $j = 1, \ldots, k$ *satisfacen las condiciones del teorema* 3A. *Entonces*

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta(H'(a))^T.$$

Si $\eta(H'(a))^T = 0$ *con probabilidad* 1, *y las matrices* H_j'', $j = 1, \ldots, k$ *existen en el punto a, entonces*

$$(H(a + b_n\eta_n) - H(a))/b_n^2 \Rightarrow \frac{1}{2}\,(\eta H_1''(a)\eta^T, \ \ldots, \ \eta H_k''(a)\eta^T).$$

Las demostraciones de estas afirmaciones, de hecho no se distinguen en nada de la demostración del teorema 3, y por eso las presentamos al lector en calidad de ejercicios. Además, proponemos convencerse de que el símbolo $\Rightarrow$ en (4)—(6) se puede sustituir por $\xrightarrow[c.s]{}$ o por $\xrightarrow[P]{}$, si se cumple $\eta_n \xrightarrow[c.s]{} \eta$ o $\eta_n \xrightarrow[P]{} \eta$, respectivamente.

El contenido de los teoremas 1—3 puede resumirse del modo siguiente. Supongamos que $\rightsquigarrow$ significa uno de los símbolos $\xrightarrow[c.s]{}$, $\xrightarrow[P]{}$, $\Rightarrow$. Entonces, si H es continua, de $\eta_n \rightsquigarrow \eta$ resulta $H(\eta_n) \rightsquigarrow H(\eta)$.

Si H es derivable en el punto a, $\eta_n \rightsquigarrow \eta$, entonces para $b_n \to 0$

$$(H(a + b_n\eta_n) - H(a))/b_n \rightsquigarrow H'(a)\eta. \qquad (7)$$

Observación 1. No es difícil notar que si a depende de n de modo que $a \equiv a_n = a_0 + o(1)$ y las derivadas en los teoremas 3, 3A y 3B son continuas, la relación (7) se conservará en la forma

$$(H(a_n + b_n\eta_n) - H(a_n))/b_n \rightsquigarrow H'(a_0)\eta. \qquad (8)$$

Para la demostración es suficiente ver que el primer miembro (8) es representable en forma de $H'(\alpha_n)\eta_n$, donde $\alpha_n = \theta a_n + (1 - \theta)(a_n + b_n\eta_n) \sim \to a_0$, $|\theta| \leqslant 1$, y utilizar el segundo teorema de continuidad.

Esa misma observación es válida para los análogos multidimensionales de la referida afirmación en los teoremas 3A, 3B.

Los teoremas enunciados conciernen a la convergencia casi segura y a la convergencia de las distribuciones. El cuarto teorema de continuidad se refiere a la convergencia de las integrales.

Teorema 4 (teorema de continuidad para los momentos). *Supongamos que* $\{\eta_n\}$ *es una sucesión de variables aleatorias numéricas y que* $\eta_n \Rightarrow \eta$ *cuando* $n \to \infty$. *En este caso, si se cumple al menos una de las condiciones siguientes*:

1) $\lim\limits_{n\to\infty} \sup \int\limits_N^\infty \mathbf{P}(|\eta_n| > x)dx \to 0$ *para* $N \to \infty$,

2) $\mathbf{P}(|\eta_n| > x) \leqslant \varphi(x)$, $\int\limits_0^\infty \varphi(x)dx < \infty$,

3) $\mathbf{M}|\eta_n|^{1+\alpha} < c < \infty$ para cierto $\alpha > 0$,

entonces $\lim\limits_{n\to\infty} \mathbf{M}\eta_n = \mathbf{M}\eta$.

Nótese que la condición 1 significa la convergencia uniforme en n hacia el cero $\int\limits_N^\infty \mathbf{P}(|\eta_n| > x)dx$ cuando $N \to \infty$.

Demostración. De la desigualdad generalizada de Chébishev,

$$\mathbf{P}(|\eta_n| > x) \leqslant \frac{\mathbf{M}|\eta_n|^{1+\alpha}}{x^{1+\alpha}},$$

se deduce que la condición 3 provoca la condición 2 y ésta, a su vez, la condición 1.

Supongamos que se ha cumplido la condición 1. Para simplificar los razonamientos, admitamos primeramente que $\eta_n \geqslant 0$. Entonces, integrando por partes, obtenemos

$$\mathbf{M}\eta_n = -\int\limits_0^\infty x\, d\mathbf{P}(\eta_n \geqslant x) = \int\limits_0^\infty \mathbf{P}(\eta_n \geqslant x)dx.$$

De esta representación, así como de la convergencia de $\mathbf{P}(\eta_n \geqslant x) \to \mathbf{P}(\eta \geqslant x)$ para casi todos los x, y de la convergencia, uniforme en n, de la integral $\int\limits_0^\infty \mathbf{P}(\eta_n \geqslant x)dx$, se deduce la legitimidad del paso límite bajo el signo de integral, en virtud del cual

$$\lim_{n\to\infty} \mathbf{M}\eta_n = \lim_{n\to\infty} \int_0^\infty \mathbf{P}(\eta_n \geqslant x)dx = \int_0^\infty \mathbf{P}(\eta \geqslant x)dx = \mathbf{M}\eta.$$

En el caso general conviene utilizar la representación $\eta_n = \eta_n^+ - \eta_n^-$, donde $\eta_n^+ = \text{máx } (\eta_n, 0)$, $\eta_n^- = \text{máx } (-\eta_n, 0)$. ◁

Señalemos que la condición 1 también puede considerarse como condición de la integrabilidad uniforme de η_n, de la cual se deduce inmediatamente la convergencia requerida de $\mathbf{M}\eta_n \to \mathbf{M}\eta$ (véase, por ejemplo, [11], [60]).

§ 6*. Función empírica de distribución como proceso aleatorio. Convergencia hacia el puente browniano

En este párrafo supondremos que se conoce el concepto de *proceso aleatorio* (digamos, en el volumen de [11]) y, en particular, las definiciones y propiedades elementales de los *procesos wieneriano y poissoniano.*

1. Distribución del proceso $nF_n^*(t)$. Nos limitaremos a examinar el caso unidimensional $\mathscr{X} = R$. Supongamos, como antes, que $F_n^*(t) = \mathbf{P}_n^*((-\infty, t))$ es la función empírica de distribución correspondiente a la muestra $X = X_n \in \mathbf{P}$.

La función $F_n^*(t)$ es una función de dos variables: t y X, o bien que es lo mismo, una función aleatoria de t o un *proceso aleatorio.*

Hallemos las *distribuciones de dimensión finita* de este proceso. Supongamos $t_1 < t_2 < \ldots < t_m$ son m puntos arbitrarios del eje numérico. Pongamos $t_0 = -\infty$, $t_{m+1} = \infty$ y designemos por

$$\Delta_j g = g(t_{j+1}) - g(t_j)$$

los incrementos de la función $g(t)$ en los semiintervalos $\Delta_j = [t_j, t_{j+1})$, $j = 0, 1, \ldots, m$. Examinemos el incremento $\Delta_j \pi_n$ del proceso

$$\pi_n(t) = nF_n^*(t).$$

Evidentemente, esto es el número de elementos de la muestra que se encuentran en Δ_j. La probabilidad de que un elemento de la muestra (digamos, x_1) se halle en Δ_j es igual a $p_j = \mathbf{P}(\Delta_j)$. Como el hecho de que los elementos tomen un valor perteneciente a Δ_j, $j = 0, 1, \ldots, m$, constituye $m + 1$ sucesos incompatibles, tenemos aquí, sin duda, una distribución polinomial (véase [11], p. 111) para el vector $(\Delta_0\pi_n, \ldots, \Delta_m\pi_n)$ con probabilidades $p_0, \ldots, p_m$, $\sum_{j=0}^{m} p_j = 1$. Como es sabido,

$$\mathbf{P}(\Delta_0\pi_n = k_0, \ldots, \Delta_m\pi_n = k_m) = \frac{n!}{k_0! \ldots k_m!} p_0^{k_0} \ldots p_m^{k_m}, \qquad (1)$$

donde $\sum_{j=0}^{m} k_j = n$.

Sea ahora $\eta(u)$, $u \in [0, 1]$, el *proceso poissoniano* continuo a la izquierda (véase [11], p. 304) *con parámetro* λ, $\eta(0) = 0$. Los incrementos de este proceso son independientes,

$$\mathbf{P}(\eta(u) = k) = e^{-\lambda u} \frac{(\lambda u)^k}{k!}.$$

Si la función de distribución $F(t) = \mathbf{P}((-\infty, t))$ es continua, podemos realizar la sustitución continua del tiempo, poniendo $u = F(t)$, $-\infty < t < \infty$, y determinar de este modo el proceso $\pi(t) = \eta(F(t))$ sobre todo el eje. Examinemos los incrementos de este proceso

$$\Delta_j \pi = \pi(t_{j+1}) - \pi(t_j) = \eta(F(t_{j+1})) - \eta(F(t_j))$$

sobre los intervalos Δ_j. Entonces

$$\mathbf{P}(\Delta_0 \pi = k_0, \ \ldots, \ \Delta_m \pi = k_m) = \prod_{j=0}^{m} e^{-\lambda p_j} \frac{(\lambda p_j)^{k_j}}{k_j!} = e^{-\lambda} \lambda^n \prod_{j=0}^{m} \frac{p_j^{k_j}}{k_j!}$$

y la probabilidad condicional de este mismo proceso, a condición de que $\pi(\infty) = \sum_{j=0}^{m} \Delta_j \pi = n$, será igual a

$$\mathbf{P}\left(\Delta_0 \pi = k_0, \ \ldots, \ \Delta_m \pi = k_m \,\middle|\, \sum_{j=0}^{m} \Delta_j \pi = n\right) =$$

$$= \frac{\mathbf{P}(\Delta_0 \pi = k_0, \ \ldots, \ \Delta_m \pi = k_m)}{\mathbf{P}(\pi(\infty) = n)} =$$

$$= \mathbf{P}(\Delta_0 \pi = k_0, \ \ldots, \ \Delta_m \pi = k_m) \frac{e^{\lambda} n!}{\lambda^n} = n! \prod_{j=0}^{m} \frac{p_j^{k_j}}{k_j!}. \quad (2)$$

Hemos obtenido para cualquier $\lambda > 0$ la misma expresión que en el segundo miembro de (1). Así pues, hemos demostrado la afirmación siguiente.

Teorema 1. *Si $F(t)$ es continua, la distribución del proceso $nF_n^*(t)$ coincide con la distribución condicional del proceso $\pi(t) = \eta(F(t))$ a condición de que $\pi(\infty) = n(\eta(1) = n)$.*

El teorema muestra que las desviaciones $n(F_n^*(t) - F(t))$ están distribuidas al igual que $\eta(F(t)) - nF(t)$ a condición de que $\eta(1) = n$ y el problema con precisión hasta la sustitución del tiempo $u = F(t)$ se reduce al estudio de las desviaciones $\eta(u) - nu$ para el proceso poissoniano condicional ($\eta(1) = n$) sobre el segmento $[0, 1]$ o bien, que es lo mismo, al estudio

de las desviaciones $n(F_n^*(t) - t)$, donde $F_n^*(t)$ corresponde a la distribución uniforme sobre $[0, 1]$.

Puede ser útil también otra representación para el proceso $nF_n^*(t)$. Sean $\zeta_1, \zeta_2, \ldots$ los puntos de saltos del proceso poissoniano $\eta(t)$, así que $\eta(\zeta_k + 0) = k$. Como es sabido ([11]), las diferencias $\xi_k = \zeta_k - \zeta_{k-1}$ $(\zeta_0 = 0)$, $k = 1, 2, \ldots$, son independientes y están distribuidas exponencialmente

$$\mathbf{P}(\xi_k > x) = e^{-\lambda x},$$

ζ_k tiene Γ-distribución con densidad (véase también el § 2.2)

$$\gamma_{\lambda,k}(x) = \frac{\lambda^k}{\Gamma_{(k)}} e^{-\lambda x} x^{k-1}.$$

Para simplificar las enunciaciones, supongamos que $F(t) = t$, $t \in [0, 1]$, $t_0 = 0$, $t_{m+1} = 1$, así que $\eta(t) = \pi(t)$.

Teorema 2. *La distribución del proceso $nF_n^*(t)$ coincide, para cualquier $v > 0$, con la distribución condicional del proceso $\pi(tv)$, $0 < t < 1$, a condición de que $\zeta_{n+1} = v$.*

Con otras palabras, la afirmación del teorema 1 seguirá válida si la condición $\pi(1) = n$ se sustituye por una condición mucho más estrecha $\pi(1) = n$, $\pi(1 + 0) = n + 1$ (suponemos que las trayectorias de $\pi(t)$ son continuas a la izquierda).

Como la probabilidad de esta nueva condición es igual a 0, puede ser que convenga añadir (véanse los §§ 4 y 8 en [11] sobre las esperanzas matemáticas, así como el § 2.9) que por distribución condicional entendemos las probabilidades

$$\mathbf{P}(A/\zeta_{n+1} = v) = \frac{\mathbf{P}(A;\ \zeta_{n+1} \in dv)}{\mathbf{P}(\zeta_{n+1} \in dv)},$$

donde $A = \{\Delta_0\pi(tv) = k_0, \ \ldots, \ \Delta_m\pi(tv) = k_m\}$, $\Delta_j\pi(tv) = \pi(t_{j+1}v) - \pi(t_j v)$.

Demostración. Representemos el suceso $\{\zeta_{n+1} \in dv\}$ en la forma del producto de dos sucesos

$$B = \{\pi(v) = n\} \text{ y } C = \{\pi(v + dv) - \pi(v) = 1\}.$$

Los sucesos B y AB no dependen de C, ya que los sucesos B y AB, por un lado, y el suceso C, por otro, se refieren a los incrementos del proceso π sobre los intervalos disjuntos del tiempo. Por eso

$$\mathbf{P}(A/\zeta_{n+1} = v) = \frac{\mathbf{P}(ABC)}{\mathbf{P}(BC)} = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} = \mathbf{P}(A/\pi(v) = n). \qquad (3)$$

Lo mismo que en (2) nos cercioramos de que esta expresión no depende de v (ni tampoco de λ) y coincide con (1). $\triangleleft$

Corolario 1. *La distribución del proceso $nF_n^*(t)$ coincide con la distribución $\pi(t\zeta_{n+1})$, $0 \leqslant t \leqslant 1$.*

Esto se deduce del hecho de que para $B = \{\Delta_0\pi(t\zeta_{n+1}) = k_0, \ldots, \Delta_m\pi(t\zeta_{n+1}) = k_m\}$ tenemos, en virtud de (3),

$$P(B) = \int \mathbf{P}(A/\zeta_{n+1} = v)\,\mathbf{P}(\zeta_{n+1} \in dv) = n! \prod_j \frac{\Delta_j^{k_j}}{k_j!}.$$

Del corolario 1 se deduce:

Corolario 2. *La distribución conjunta de los elementos de la serie variacional $x_{(1)}, \ldots, x_{(n)}$ de la muestra X de la distribución uniforme coincide con la distribución conjunta*

$$\frac{\zeta_1}{\zeta_{n+1}}, \ldots, \frac{\zeta_n}{\zeta_{n+1}},$$

o bien, que es lo mismo, la distribución conjunta de las diferencias $x_{(1)}$, $x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(n-1)}$, $1 - x_{(n)}$ coincide con la distribución conjunta

$$\frac{\xi_1}{\zeta_{n+1}}, \ldots, \frac{\xi_{n+1}}{\zeta_{n+1}}.$$

Para concluir este apartado determinaremos los momentos de segundo orden para los incrementos del proceso $n(F_n^*(t) - F(t))$. Para nosotros será más cómodo examinar el proceso

$$w^n(t) = \sqrt{n}(F_n^*(t) - F(t)).$$

Es evidente que $\mathbf{M}\Delta_j w^n = 0$, $\mathbf{M}(\Delta_j w^n)^2 = \Delta_j F(1 - \Delta_j F)$. Para calcular los momentos mixtos notemos que $(i \neq j)$

$$\mathbf{M}(\Delta_i w^n \cdot \Delta_j w^n) = \frac{1}{n} \sum_{k,l=1}^{n} \mathbf{M}(\mathbf{I}_{x_k}(\Delta_i) - \mathbf{P}(\Delta_i)) \times$$

$$\times (\mathbf{I}_{x_l}(\Delta_j) - \mathbf{P}(\Delta_j)) = \frac{1}{n} \sum_{k,\ l=1}^{n} \{\mathbf{M}\mathbf{I}_{x_k}(\Delta_i)\mathbf{I}_{x_l}(\Delta_j) - \mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j)\}.$$

Puesto que

$$\mathbf{M}\mathbf{I}_{x_k}(\Delta_i)\mathbf{I}_{x_l}(\Delta_j) = \begin{cases} \mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j) \text{ para } k \neq l \\ 0 \text{ para } k = l, \end{cases}$$

Entonces $\mathbf{M}(\Delta_i w^n \cdot \Delta_j w^n) = -\mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j) = -\Delta_i F \cdot \Delta_j F$.

Ahora bien, los incrementos del proceso w^n están correlacionados negativamente.

2. Comportamiento límite del proceso $w^n(t)$. Supongamos que $F(t)$ es continua. Del punto 1 entonces se deduce que podemos limitarnos a examinar la distribución $F(t) = t$ uniforme sobre $[0, 1]$, $0 \leqslant t \leqslant 1$.

Designemos por $w(t)$ el *proceso wieneriano estándar*, o sea, el proceso con incrementos independientes para el cual $w(t)$ está distribuido normalmente con parámetros $(0, t)$. El proceso

$$w^\circ(t) = w(t) - tw(1)$$

se llama *puente browniano* (puesto que en él se hallan asegurados ambos extremos: $w^\circ(0) = w^\circ(1) = 0$). La distribución de este proceso coincide con la distribución condicional del proceso $w(t)$ a condición de que $w(1) = 0$ (mejor dicho, es necesario adoptar la condición $|w(1)| < \varepsilon$ y pasar al límite para $\varepsilon \to 0$).

Resulta que las distribuciones de dimensión finita de los procesos

$$w^n(t) = \sqrt{n}(F_n^*(t) - F(t)), \quad t \in [0, 1],$$

convergen, cuando $n \to \infty$, hacia las distribuciones correspondientes del puente browniano $w^\circ(t)$.

Este hecho permite aproximar los procesos $w^n(t)$, llamados, a veces, *procesos empíricos*, con ayuda del proceso $w^\circ(t)$. Precisamente por eso podemos imaginarnos que, con grandes valores de n, tiene lugar la igualdad aproximada

$$\sqrt{n}(F_n^*(t) - F(t)) \approx w^\circ(t) \tag{4}$$

que describe la distribución de las desviaciones de $F_n(t)$ respecto a $F(t)$ (recordemos que aquí hemos considerado que $F(t) = t$, $t \in [0, 1]$.

No obstante, necesitaremos la afirmación del tipo (4) en una forma más fuerte. Examinemos, por ejemplo, la estadística $U = \sqrt{n} \sup_t (F_n^*(t) - F(t))$. Dicha afirmación hace natural la suposición de que con grandes valores de n la variable aleatoria U está distribuida aproximadamente al igual que $\sup_{0 \leqslant t \leqslant 1} w^\circ(t)$. Pero de nuestra afirmación esto no se deduce de ningún modo, puesto que U no puede ser representada como función de los valores de $w^n(t) = \sqrt{n}(F_n^*(t) - F(t))$ en cualquier número finito de puntos. Por eso es mucho más fuerte la siguiente afirmación.

Designemos por $D(a, b)$ el espacio de las funciones sobre el segmento $[a, b]$, que son continuas a la izquierda (en el punto a a la derecha) y tienen sólo un número finito de saltos, y designemos por $C(a, b)$ el espacio de todas las funciones continuas sobre $[a, b]$. Es evidente que la trayectoria

$w^n(t)$ pertenece a $D(0, 1)$. Además, es sabido (véase [11], capítulo 13) que las trayectorias $w^\circ(t)$ pertenecen a $C(0, 1)$ con probabilidad 1. Para simplificar la exposición podemos suponer que todas las trayectorias $w(t)$ y, por consiguiente, $w^\circ(t)$ se encuentran en $C(0, 1)$ (véase [11]). Como $C(0, 1) \subset D(0, 1)$, entonces $(D(0, 1), \sigma_D)$ — donde σ_D es el σ-álgebra de los subconjuntos de $D(0, 1)$, engendrada por conjuntos cilíndricos *) — puede ser considerado como el espacio muestral **) de los procesos w^n y w°.

Teorema 3 (teorema funcional del límite para los procesos empíricos). *Sea f la funcional que está definida sobre el espacio $D(0, 1)$ y que posee las propiedades siguientes:*

1) *$f(w_n)$ y $f(w^\circ)$ son magnitudes aleatorias (o sea, $f(y)$ realiza la aplicación medible $(D(0, 1), \sigma_D)$ en $(R, \mathfrak{B})$);*

2) *$f(y)$ es una funcional que es continua en los "puntos" del espacio $C(0, 1)$ con respecto a la métrica uniforme, o sea, $f(y_n) \to f(y)$ para $n \to \infty$ si $y \in C(0, 1)$ y $\varrho(y_n, y) = \sup_{0 \leqslant t \leqslant 1} |y_n(t) - y(t)| \to 0$.*

Si estas condiciones han sido cumplidas, entonces

$$f(w^n) \Rightarrow f(w^\circ).$$

Si la funcional f es continua en la métrica uniforme en todo punto $y \in D(0, 1)$, la condición 1) se cumple automáticamente.

Es evidente que la funcional U, examinada anteriormente, satisface las condiciones del teorema, así que para $n \to \infty$,

$$U \Rightarrow \sup_{0 \leqslant t \leqslant 1} w^\circ(t).$$

Como en esta relación, la distribución del segundo miembro se puede hallar en forma explícita (véase, por ejemplo, [5], [58]):

$$\mathbf{P}\left(\sup_{0 \leqslant t \leqslant 1} w^\circ(t) > z\right) = e^{-2z^2},$$

obtenemos, de este modo, la expresión aproximada para la distribución de U.

El uso del teorema 3 para el cálculo de la distribución límite de otras estadísticas se examina en los párrafos siguientes.

La demostración del teorema 3 se da en el Suplemento II.

---

*) O sea, por los conjuntos que tienen la forma $\{y(t_1) \in B_1, \ldots, y(t_m) \in B_m\}$, donde $B_1, \ldots, B_m$ son los conjuntos de Borel.

**) (D_0, σ) es el espacio muestral del proceso $\xi(t)$ si en él está dada la distribución del conjunto ξ de tal modo que las trayectorias $\xi(t)$ se encuentran en D_0.

§ 7. Distribución límite para las estadísticas de primer tipo

Recordemos que llamamos estadísticas de primer tipo las estadísticas $S_n(X) = G(F_n^*)$, donde la funcional G tiene la forma $G(F) = h\left(\int g(x)\, dF(x)\right)$. Con otras palabras,

$$S_n(X) = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right).$$

Ya hemos visto (teorema 3.1) que si $X \in F_0$ y h es continua en el punto $a = \int g(x)\, dF_0(x)$, entonces $S_n \underset{c.s}{\longrightarrow} h(a)$.

Teorema 1. *Si $X \in F_0$, h es derivable en el punto a, $\int g^2(x)\, dF_0(x) < \infty$, entonces*

$$\sqrt{n}(S_n(X) - h(a)) \Rightarrow h'(a)\xi,$$

donde $\xi \in \Phi_{0,\sigma^2}$, $\sigma^2 = \int (g(x) - a)^2 dF_0(x)$. Φ_{0,σ^2} aquí significa la distribución normal con parámetros $(0, \sigma^2)$.

Demostración. Representemos la estadística $S_n(X)$ en la forma

$$h\left(a + \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - a)\right]\right),$$

donde, según el teorema central del límite (véase [11]),

$$\eta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - a) \in \Phi_{0,\sigma^2},$$

$$\sigma^2 = \mathbf{M}(g(x_1) - a)^2 = \int (g(x) - a)^2 dF_0(x).$$

Nos queda hacer uso del tercer teorema de continuidad para $b_n = 1/\sqrt{n}$. ◁

A veces es más cómodo examinar las funcionales de primer tipo en la forma $G(F) = h\left(\int g(x) d(F - F_0)\right)$. Evidentemente, todo lo dicho también es válido para éstas, con la única diferencia de que a ha de considerarse igual a 0.

Citemos el análogo del teorema 1 para el caso en que la función $g = (g_1, \ldots, g_s)$ es el vector (o sea $G(F) = h\left(\int g_1(x) dF(x), \ldots, \int g_s(x) dF(x)\right)$.

Teorema 1A. *Supongamos que $S_n(X) = G(F_n^*)$, $h(t)$ es derivable en el*

punto $a = \int g(x)dF_0(x)$, *y que la matriz de los segundos momentos* $\sigma^2 = \|\sigma_{ij}\| = \mathbf{M}(g(x_1) - a)^T(g(x_1) - a)$ *es finita. Entonces*

$$(S_n(X) - h(a))\sqrt{n} \Rightarrow \xi(h'(a))^T = \sum_{j=1}^{s} \frac{\partial h(a)}{\partial t_j}\,\xi_j, \tag{1}$$

donde $\xi = (\xi_1, \ldots, \xi_s) \in \Phi_{0,\sigma^2}$.

Si $\xi(h'(a))^T = 0$ *con probabilidad* 1, *y la matriz de segundas derivadas* $h''(t) = \left\| \frac{\partial^2}{\partial t_i \partial t_j} h(t) \right\|$ *existe en el punto a, entonces*

$$(S_n(X) - h(a))n \Rightarrow \frac{1}{2}\,\xi h''(a)\xi^T = \frac{1}{2} \sum_{i,j=1}^{s} \frac{\partial^2 h(t)}{\partial t_i \partial t_j}\,\xi_i \xi_j.$$

Para la demostración del teorema 1A conviene usar el teorema de continuidad 5.3A y el teorema central del límite multidimensional, en virtud del cual $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (g(x_i) - a) \Rightarrow \xi$ (véase el suplemento V).

Completamente análogo es el teorema de la distribución límite $S_n(X)$ cuando la función h, y junto con ella también la estadística $S_n(X)$, son vectores. El lector reproducirá sin dificultad su enunciación y demostración con ayuda del teorema 5.3B.

Ejemplo 1. Supongamos que $X \in \mathbf{P}_0$ y $\mathbf{P}_0$ es tal que $\mathbf{M}x_1 = \alpha > 0$, $\mathbf{D}x_1 = d^2 < \infty$. ¿Qué representa en estas condiciones la distribución límite de la estadística $S = 1/\bar{x}\left(\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$? Aquí, las condiciones del teorema 1 están evidentemente cumplidas para $h(t) = 1/t$, $g(x) = x$, con la particularidad de que $a = \alpha$, $\sigma^2 = d^2$, $h(a) = 1/\alpha$, $h'(a) = -1/\alpha^2$. En virtud del teorema 1,

$$(S - 1/\alpha)\sqrt{n} \Rightarrow -\xi/\alpha^2, \quad \xi \in \Phi_{0,d^2},$$

así que

$$(S - 1/\alpha)\sqrt{n} \in \Phi_{0,d^2/\alpha^4}.$$

Ejemplo 2. Hallemos la distribución límite de la estadística

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

si $\mathbf{M}x_1 = \alpha$, $\mathbf{D}x_i = d^2$ y $\mathbf{M}x_1^4 < \infty$. (Ya sabemos que en virtud del primer

teorema de continuidad, $S^2 \xrightarrow[cs]{} d^2$). No es difícil hallar directamente la distribución límite necesaria, utilizando las representaciones

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

$$(S^2 - d^2)\sqrt{n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(x_i - \alpha)^2 - d^2] - \sqrt{n}(\bar{x} - \alpha)^2.$$

No obstante, haremos uso del teorema 1A. Según los datos de este teorema debemos suponer que

$$G(F) = \int (x - \alpha)^2 dF(x) - \left(\int x\, dF(x) - \alpha\right)^2,$$

así que $g_1(x) = (x - \alpha)^2$, $g_2(x) = x$, $h(t) = t_1 - (t_2 - \alpha)^2$. Puesto que en el punto $a = (d^2, \alpha)$

$$\frac{\partial h(a)}{\partial t_1} = 1, \quad \frac{\partial h(a)}{\partial t_2} = 0,$$

entonces

$$(S^2 - d^2)\sqrt{n} \Rightarrow \xi, \quad \xi \in \Phi_{0,v^2}; \quad v^2 = \mathbf{M}(x_1 - \alpha)^4 - d^4.$$

Ejemplo 3. *Estadística* χ^2. Concluyendo este párrafo examinemos un ejemplo de estadística que puede pertenecer tanto a la del tipo I como a la del tipo II.

Examinemos las estadísticas construidas con ayuda de la funcional que tiene la forma

$$G(F) = h\left(\int g\, dF\right). \tag{2}$$

donde g es la función de variación limitada sobre el segmento $[a, b]$ tal que $F(a) = 0$, $F(b) = 1$ (a y b pueden ser infinitos). Como $\int g\, dF = g(b) - \int F\, dg$, la funcional $G(F)$ será continua en la métrica uniforme si sólo es continua la función h. Es fácil comprender que la clase destacada de características no es sino la intersección de las clases de estadísticas de los tipos I y II.

Lo mismo es válido en el caso en que g es una función de forma vectorial con componentes g_i que tienen una variación limitada.

Examinemos ahora la partición del eje real (espacio $\mathcal{X}$) en los intervalos disjuntos $\Delta_1, \ldots, \Delta_r$, y designemos $\nu_i = n\mathbf{P}_n^*(\Delta_i)$, $p_i = \mathbf{P}_0(\Delta_i)$ ($\mathbf{P}_0$ es la distribución correspondiente a F_0, así que $X \in \mathbf{P}_0$). Se llama estadística "ji-

cuadrado" $\chi^2 = \chi^2(X)$ la estadística

$$\chi^2(X) = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

Evidentemente que esto es una estadística de tipo II, ya que ella corresponde, con una exactitud de hasta el factor n, a la funcional

$$G(F) = G_1(P) = \sum_{i=1}^{r} \frac{(\mathbf{P}(\Delta_i) - \mathbf{P}_0(\Delta_i))^2}{\mathbf{P}_0(\Delta_i)}.$$

Para representar $\chi^2(X)$ como estadística de tipo I, examinemos la funcional que tiene la forma (2)

$$G(F) = h\left(\int g d(F - F_0)\right)$$

con la función $h(u) = \sum_{j=1}^{r} u_j^2$ y la función vectorial g con coordenadas

$$g_j(x) = \begin{cases} 1/\sqrt{p_j} & \text{para } x \in \Delta_j, \\ 0 & \text{para } x \notin \Delta_j. \end{cases}$$

Como la función h es derivable, $\frac{\partial h(0)}{\partial u_j} = 0$, $\frac{\partial^2 h(0)}{\partial u_i \partial u_j} = 2\delta_{ij}$ (δ_{ij} es el símbolo de Kronecker), entonces, poniendo $S_n(X) = G(F_n^*)$, obtenemos

$$nS_n(X) = n \sum_{j=1}^{r} \left[\left(\frac{\nu_j}{n} - p_j \right) \frac{1}{\sqrt{p_j}} \right]^2 = \chi^2(X).$$

Para $X \in P_0$, en virtud de la segunda parte del teorema 1A,

$$\chi^2(X) \Rightarrow \sum_{j=1}^{r} \xi_j^2, \tag{3}$$

donde $\xi = (\xi_1, \ldots, \xi_r)$ es el vector normalmente distribuido (límite para $\left(\frac{\nu_1 - np_1}{\sqrt{np_1}}, \ldots, \frac{\nu_r - np_r}{\sqrt{np_r}} \right)$) con la media nula y la matriz $\sigma^2 = \|\sigma_{ij}\|$ de segundos momentos

$$\sigma_{ij} = \mathbf{M}\xi_i\xi_j = \mathbf{M}(g_i(x_1) - \sqrt{p_i})(g_j(x_1) - \sqrt{p_j})$$

(de la definición de g_j se deduce que $\mathbf{M}g_j(x_1) = \sqrt{p_j}$). Puesto que $g_i(x)g_j(x) = 0$ para $i \neq j$ y $\mathbf{P}(g_j^2(x_1) = 1/p_j) = p_j$, $\mathbf{P}(g_j^2(x_1) = 0) = 1 - p_j$, entonces

$$\sigma_{ij} = \delta_{ij} - \sqrt{p_i p_j}.$$

Aclaremos ahora qué representa la distribución del segundo miembro en (3) (o sea, la distribución límite $\chi^2(X)$).

Examinemos la transformación ortogonal en R^r con la matriz C y examinemos el vector

$$\eta = \xi C.$$

El vector η, al igual que ξ, será distribuido normalmente. En efecto, la normalidad de la magnitud ξ quiere decir que su función característica es igual a (véase [11])

$$\mathbf{M}e^{it\xi^T} = e^{-\frac{1}{2}t\sigma^2 t^T},$$

donde $\sigma^2 = \|\sigma_{ij}\|$ es la matriz de segundos momentos. Pero f.c. para η

$$\mathbf{M}e^{it\eta^T} = \mathbf{M}e^{itC^T\xi^T} = e^{-\frac{1}{2}tC^T\sigma^2 Ct^T}$$

tiene la misma forma y, por consiguiente, η es un vector normal, pero con la matriz de segundos momentos $d^2 = C^T\sigma^2 C = \|d_{ij}\|$, así que

$$d_{ij} = \mathbf{M}\eta_i\eta_j = \sum_{k,l} c_{li}\sigma_{lk}c_{kj} = \sum_{k,l} c_{li}(\delta_{lk} - \sqrt{p_l p_k})c_{kj} =$$

$$= \sum_l c_{li}c_{lj} - \left(\sum_l c_{li}\sqrt{p_l}\right)\left(\sum_k c_{kj}\sqrt{p_k}\right). \qquad (4)$$

Escojamos ahora la matriz C de modo que su primera columna tenga las coordenadas $c_{l1} = \sqrt{p_l}$ (esto corresponde a la fijación del primer vector del sistema transformado de las coordenadas y es posible, ya que $\sum_{l=1}^{r} c_{l1}^2 = \sum p_l = 1$). En este caso es evidente que el segundo sumando en (4), en virtud de la ortogonalidad de C, es igual a 1 sólo para $i = j = 1$, y es igual a 0 en el caso contrario. Esto significa que $d_{11} = \mathbf{M}\eta_1^2 = 0$, $d_{ij} = \mathbf{M}\eta_i\eta_j = \delta_{ij}$ para $i \geqslant 2$, y por consiguiente, $\eta_1 = 0$ con una probabilidad igual a 1, y las magnitudes $\eta_2, \ldots, \eta_r$ son independientes y están distribuidas normalmente con los parámetros (0, 1). A base de la ortogonalidad de C obtenemos

$$\sum_{j=1}^{r} \xi_j^2 = \sum_{j=1}^{r} \eta_j^2 = \sum_{j=2}^{r} \eta_j^2,$$

$$\chi^2(X) \Rightarrow \sum_{j=2}^{r} \eta_j^2. \qquad (5)$$

En esta igualdad, la distribución del segundo miembro se llama *distribución* χ^2 ("ji-cuadrado") con $r - 1$ grados de libertad (véase [11] y también

el § 2.2). En la exposición ulterior encontraremos muchas veces esta distribución.

Una demostración más de (5) será obtenida en el párrafo siguiente. Además, (5) será demostrado en el § 3.16 con ayuda de consideraciones más generales.

Algunos otros ejemplos de uso de los teoremas 1 y 1A se dan en los capítulos posteriores.

§ 8. Distribución límite para las estadísticas de segundo tipo

Aquí nos limitaremos a examinar el caso $\mathscr{X} = R$. La funcional $G(F_n^*)$ sujeta a estudio será una magnitud aleatoria si ella realiza la aplicación medible $(D(-\infty, \infty), \sigma_D)$ en $(R, \mathfrak{B})$. Sin embargo, en lo sucesivo nos será más cómodo estudiar las funcionales que no están definidas sobre $D(-\infty, \infty)$ sino sobre $D(0, 1)$ (compárense con el § 6).

Para hacer esto apliquemos $D(-\infty, \infty)$ en $D(0, 1)$. Supongamos que la función de distribución F_0, correspondiente a la muestra, es continua y monótona, así que está definida la función inversa $F_0^{-1}(t)$ (igual a la cuantila de orden t de F_0). Nos será suficiente examinar los valores de $G(F)$ para las funciones F, cuyo portador está presente en el portador de F_0. A cada F pongámosle en correspondencia la función

$$\tilde{F}(t) = F(F_0^{-1}(t)) \equiv FF_0^{-1}(t).$$

Es evidente que $N_{\tilde{F}} \subseteq [0, 1]$, donde $N_{\tilde{F}}$ es el portador de $\tilde{F}$, así que $\tilde{F} \in D(0, 1)$ es precisamente la función de distribución. La transformación inversa de $D(0, 1)$ en $D(-\infty, \infty)$ se lleva a cabo por la igualdad

$$F(u) = \tilde{F}(F_0(u)) \equiv \tilde{F}F_0(u).$$

Pongamos ahora en correspondencia con la funcional G la funcional $\tilde{G}$ definida sobre las funciones de distribución $H \in D(0, 1)$ ($N_H \subseteq [0, 1)$ por la igualdad

$$\tilde{G}(H) = G(HF_0). \tag{1}$$

La inversión de esta fórmula tiene la forma

$$G(F) = \tilde{G}(FF_0^{-1}).$$

Estas igualdades reducen el estudio de las funcionales $G(F)$ al estudio de las funcionales $\tilde{G}(H)$ definidas en las funciones de distribución de $D(0, 1)$. En virtud de estas igualdades,

$$G(F_n^*) = \tilde{G}(F_n^*F_0^{-1}) = \tilde{G}(D_n^*). \tag{2}$$

$$D_n^* = F_n^*F_0^{-1} \tag{3}$$

no es otra cosa sino la función empírica de distribución de la muestra desde la distribución uniforme sobre [0, 1]. En efecto, según el teorema 6.1, el proceso $nD_n^*(t) = nF_n^*(F_0^{-1}(t))$ tiene la misma distribución que el proceso poissoniano $\pi(F_0(F_0^{-1}(t))) = \pi(t)$, $t \in [0, 1]$ (con un parámetro $\lambda > 0$) a condición de que $\pi(1) = n$. En virtud de ese mismo teorema 6.1, esto demuestra la afirmación requerida.

Lo dicho significa que el estudio de $G(F_n^*)$ se reduce a la investigación de la funcional $\tilde{G}$ de la distribución empírica que corresponde a la distribución uniforme sobre [0, 1].

Ejemplo 1. Sea $G(F) = \zeta_p$ la cuantila de orden p de la función de distribución F. Entonces $\tilde{G}(H) = G(HF_0)$ será la cuantila de orden p de la función de distribución HF_0 o bien, que es lo mismo (supongamos, para simplificar, que H es continua), la solución de la ecuación $H(F_0(t)) = p$, igual a $F_0^{-1}(H^{-1}(p))$.

Esto significa que la cuantila muestral $\zeta_p^* = G(F_n^*) = \tilde{G}(D_n^*)$ (véanse (2) y (3)) de la muestra $X \in F_0$ no es otra cosa sino el valor de la función F_0^{-1} de la cuantila muestral $\eta_p^* = (D_n^*)^{-1}(p)$ de orden p de la muestra Y de la distribución uniforme.

Por lo tanto, si logramos hallar la distribución límite de η_p^*, entonces la distribución límite de ζ_p^* podrá ser obtenida con ayuda de los teoremas de continuidad.

Ejemplo 2. Examinemos la funcional $G(F) = \sup_{-\infty < t < \infty} |F(t) - F_0(t)|$. En este caso

$$\tilde{G}(H) = G(HF_0) = \sup_{-\infty < t < \infty} |H(F_0(t)) - F_0(t)| = \sup_{u \in [0,1]} |H(u) - u|,$$

así que

$$G(F_n^*) = G(D_n^*) = \sup_{u \in [0,1]} |D_n^*(u) - u|,$$

y en correspondencia con el contenido del § 6, la distribución de la estadística $G(F_n^*)$ no dependerá de F_0 si F_0 es continua. En este sentido la estadística $G(F_n^*)$ puede llamarse invariante respecto a la distribución uniforme de la muestra.

Ejemplo 3. La funcional

$$G(F) = \int_{-\infty}^{\infty} |F(t) - F_0(t)|^k dF_0(t)$$

también engendra la estadística $G(F_n^*)$, invariante respecto a F_0, ya que

$$\tilde{G}(H) = \int_0^1 |H(u) - u|^k du, \qquad G(F_n^*) = \int_0^1 |D_n^*(u) - u|^k du.$$

Ejemplo 4. Examinemos la funcional

$$G(F) = \sum_{j=1}^{r} \frac{(\Delta_j F - \Delta_j F_0)^2}{\Delta_j F_0},$$

donde $\Delta_j F$ son los incrementos de la función F sobre los intervalos $\Delta_j = [t_j, t_{j+1})$ que forman la partición de una recta real. Evidentemente que $nG(F_n^*)$ no es otra cosa sino la estadística χ^2 examinada en el ejemplo 7.3 en calidad de estadística de tipo I.

Tenemos

$$\bar{G}(H) = G(HF_0) = \sum_{j=1}^{r} \frac{(\Delta_j HF_0 - \Delta_j F_0)^2}{\Delta_j F_0},$$

donde

$$\Delta_j HF_0 = H(F_0(t_{j+1})) - H(F_0(t_j)) = \delta_j H,$$

$\delta_j H$ son los incrementos de H sobre los intervalos $\delta_j = [\tau_j, \tau_{j+1})$, $\tau_j = F_0(t_j)$. Así, pues, designando con esa misma letra δ_j la longitud del intervalo δ_j, obtenemos

$$G(F_n^*) = \bar{G}(F_n^* F_0) = \bar{G}(D_n^*) = \sum_{j=1}^{r} (\delta_j D_n^* - \delta_j)^2/\delta_j.$$

Aquí el segundo miembro es la estadística $n^{-1}\chi^2$ para la muestra Y de la distribución uniforme con partición $\{\delta_j\}$. Esto significa, en particular, que en el ejemplo 3 del párrafo precedente pudiéramos limitarnos a examinar la distribución uniforme F_0, aunque la estadística χ^2 por sí misma no es invariante con respecto a F_0.

Ahora bien, podemos, sin limitar la generalidad, suponer que la funcional $G(F)$ se da sobre $D(0, 1)$ y $F_0(t) = t$, $t \in [0, 1]$. El paso a las funcionales "iniciales" se realiza mediante las fórmulas (1) y (2) y será ilustrado con otros ejemplos.

Con el fin de encontrar la distribución límite para las funcionales de segundo tipo $G(F_n^*)$ es necesario, al igual que en el apartado precedente, imponer a las funcionales ciertas condiciones de suavidad.

Pongamos para abreviar, $\|x\| = \sup_{0 \leqslant t \leqslant 1} |x(t)|$.

Definición 1. La funcional $G(F)$ se llama *continuamente derivable de orden k en el punto* F_0 si existe la funcional $g(F_0, v)$ que para cualquier función $v \in C(0, 1)$ y cualquier sucesión $v_h \in D(0, 1)$ es tal que $\|v_h - v\| \to 0$ cuando $h \to 0$ satisface las relaciones

$$\frac{G(F_0 + hv_h) - G(F_0)}{h^k} \to g(F_0, v), \tag{4}$$

$$g(F_0, v_h) \to g(F_0, v).$$

La última relación significa, evidentemente, la continuidad en la métrica uniforme en los puntos de $C(0, 1)$ de la funcional $g(F_0, v)$ que se puede llamar *derivada de orden k de G en la dirección de v*.

Observación 1. Recordemos que aquí, en cualquier parte, por F_0 se puede entender la distribución uniforme sobre [0, 1].

Mostremos que en el ejemplo 1, la funcional $G(F) = F^{-1}(p)$ de la distribución F sobre [0, 1] es continuamente derivable en el "punto" $F_0(t) = t$, $t \in [0, 1]$.

En efecto, por definición,

$$G(F_0 + hv_h) = \text{máx}\,\{t\colon F_0(t) + hv_h(t) \leqslant p\}.$$

Como esta funcional es continua en la métrica uniforme en el punto F_0, podemos poner $G(F_0 + hv_h) = p + \delta$, donde $\delta = \delta(h) \to 0$ para $h \to 0$. Luego, de la relación $\|v_h - v\| \to 0$, donde $v \in C(0, 1)$, se deduce $|v_h(p + \delta) - v_h(p)| \equiv r(h) \to 0$ cuando $h \to 0$. Como $F_0(p + \delta) = p + \delta$, para $t = G(F_0 + pv_h) = p + \delta$ obtenemos

$$F_0(t) + hv_h(t) = p + \delta + hv_h(p + \delta) = p + \delta + h(v_h(p) + \tau r(h)) \leqslant p,$$

donde $|\tau| \leqslant 1$. La igualdad inversa análoga se puede escribir valiéndose del hecho de que $F_0(t + 0) + hv_h(t + 0) \geqslant p$. De aquí se deduce que $\delta = -h(v_h(p) + \tau_1 r(h))$, $|\tau_1| \leqslant 1$, así que

$$\frac{G(F_0 + hv) - G(F_0)}{h} = \frac{\delta}{h} \to -v(p).$$

Ahora bien, la derivada $g(F_0, v)$ en este ejemplo es igual a

$$g(F_0, v) = -v(p). \quad \triangleleft \tag{5}$$

Es evidente que en el ejemplo 2, la funcional $G(F) = \sup_{t \in [0,1]} |F(t) - F_0(t)|$ es también continuamente derivable en toda dirección, ya que $G(F_0) = 0$,

$$g(F_0, v) = \frac{G(F_0 + hv)}{h} = \sup_{t \in [0,1]} |v(t)|.$$

En el ejemplo 3, la funcional $G(F) = \int_0^1 |F(t) - F_0(t)|^k dR(t)$ para cualquier función de variación limitada $R(t)$ es continuamente derivable (de orden k) en toda dirección, ya que

$$g(F_0, v) = \frac{G(F_0 + hv)}{h^k} = \int_0^1 |v(t)|^k dR(t).$$

La afirmación análoga es válida respecto al ejemplo 4 sobre la funcional

$$G(F) = \sum_{j=1}^{r} \frac{(\Delta_j F - \Delta_j F_0)^2}{\Delta_j F_0}$$

la cual será continuamente derivable de segundo orden, puesto que para ella

$$g(F_0, v) = \frac{G(F_0 + hv)}{h^2} = \sum_{j=1}^{r} \frac{(\Delta_j v)^2}{\Delta_j F_0}.$$

En los ejemplos 2 — 4, la generalización de las funcionales son las funcionales de forma $G(F) = G_1(F - F_0)$, donde la funcional G_1 es homogénea en el sentido de que $G_1(hv) = h^k G(v)$. Es evidente que todas estas funcionales serán derivables.

Enunciemos ahora el teorema principal de las funcionales de segundo tipo. Sea, como antes, $F_0(t) \equiv t$, $t \in [0, 1]$.

Teorema 1. *Si $X \in F_0$ y la funcional $G(F)$ es derivable (de orden k) en sentido de la definición* 1, *entonces*

$$[G(F_n^*) - G(F_0)]n^{k/2} \Rightarrow g(F_0, w^\circ),$$

donde w° es el puente browniano.

Demostración. Es sabido (véase, por ejemplo, [5]) que los compactos en el espacio métrico de las funciones continuas $C(0, 1)$ con métrica uniforme, se describen del modo siguiente. A cada función $\varphi(\Delta) > 0$, $\varphi(\Delta) \to 0$ para $\Delta \to 0$, y al número $N > 0$ le corresponde el compacto

$$K = K(\varphi, N) = \{y \in C(0, 1): \omega_\Delta(y) \leqslant \varphi(\Delta), |y(0)| \leqslant N\},$$

donde $\omega_\Delta(y)$ es el módulo de continuidad y:

$$\omega_\Delta(y) = \sup_{|t-u| \leqslant \Delta} |y(t) - y(u)|.$$

Designemos por K_h el conjunto

$$K_h = \{y \in D(0, 1): \omega_\Delta(y) \leqslant \varphi(\Delta) \text{ para todos } \Delta \geqslant h: |y(0)| \leqslant N\}.$$

Los conjuntos K_h podrían llamarse "precompactos" (este término se utiliza en el análisis funcional en otro sentido) engendrados por el compacto K. Está claro que $K_{h_1} \subset K_{h_2}$ para $h_1 \leqslant h_2$, $\bigcap_{n=1}^{\infty} K_{1/n} = K$ y que $K_h \subset (K)^{\varphi(h)}$, donde $(K)^\varepsilon$ es el ε-entorno del conjunto K.

Mostremos ahora que para $\delta > 0$ dado existe el compacto K (y, por lo tanto, la familia de los precompactos K_h que le corresponden) y la suce-

sión $h_n \to 0$ para $n \to \infty$ tales que

$$\lim_{n\to\infty} \sup \mathbf{P}(w^n \notin K_{h_n}) \leqslant \delta. \tag{6}$$

En efecto, según el teorema 6.3, para toda funcional f que sea continua en la métrica uniforme se cumple $f(w^n) \Rightarrow f(w^\circ)$, donde $w^n(t) = \sqrt{n}(F_n^*(t) - t)$, $0 \leqslant t \leqslant 1$. Como $\omega_\Delta(y)$ es tal funcional, entonces $\omega_\Delta(w^n) \Rightarrow \omega_\Delta(w^\circ)$. Pero $\omega_\Delta(w^\circ) \underset{\text{c.s}}{\to} 0$ para $\Delta \to 0$, ya que las trayectorias de w° son casi seguramente continuas. Por consiguiente, para ε y δ dados, siendo Δ suficientemente pequeño,

$$\mathbf{P}(\omega_\Delta(w^\circ) > \varepsilon) \leqslant \delta.$$

Considerando, sin limitar la generalidad, el número ε como punto de continuidad de la distribución $\omega_\Delta(w^\circ)$, obtenemos

$$\lim_{n\to\infty} \sup \mathbf{P}(\omega_\Delta(w^n) > \varepsilon) \leqslant \delta.$$

Sea ahora $\varepsilon_k \downarrow 0$ cierta sucesión, y los números $\Delta_k \downarrow 0$ son tales que

$$\lim_{n\to\infty} \sup \mathbf{P}(\omega_{\Delta_k}(w^n) > \varepsilon_k) \leqslant \delta/2^{k+1}.$$

Formemos la función $\varphi(\Delta) = \varepsilon_k$ para $\Delta \in [\Delta_{k+1}, \Delta_k)$. Es evidente que $\varphi(\Delta) \to 0$ para $\Delta \to 0$, y podemos examinar los precompactos K_h construidos según la función φ. Entonces para todo $k < \infty$,

$$\lim_{n\to\infty} \sup \mathbf{P}(w^n \notin K_{\Delta_k}) \leqslant \lim_{n\to\infty} \sup \sum_{j=1}^{k+1} \mathbf{P}(\omega_{\Delta_j}(w^n) > \varepsilon_j) \leqslant$$

$$\leqslant \sum_{j=1}^{k+1} \lim_{n\to\infty} \sup \mathbf{P}(\omega_{\Delta_j}(w^n) > \varepsilon_j) \leqslant \delta/2$$

(para $k = \infty$ esta desigualdad puede ser injusta). La relación obtenida quiere decir que para cada δ existe la sucesión $h_n \to 0$ cuando $n \to \infty$ es tal que se cumple (6). Examinemos ahora la magnitud

$$[G(F_n^*) - G(F_0)]n^{k/2} = g(F_0, w^n) + H_n(w^n),$$

donde $H_n(x) = [G(F_0 + x/\sqrt{n}) - G(F_0)]n^{k/2} - g(F_0, x)$. Puesto que, en virtud del teorema 6.3 y la definición 1, $g(F_0, w^n) \Rightarrow g(F_0, w^\circ)$, basta con que nos cercioremos de que

$$H_n(w^n) \underset{P}{\to} 0. \tag{7}$$

Nótese que para todo compacto $K \subset C(0, 1)$ y para toda sucesión $h_n \to 0$ cuando $n \to \infty$,

$$\sup_{\substack{x\in D(0,1)\\ x\in(K)^h_\infty}} |H_n(x)| \to 0.$$

Admitiendo lo contrario, llegaremos a la existencia de una sucesión $x_n \in D(0, 1)$ tal que $|x_n - x| \to 0$, $x \in C(0, 1)$, $\lim\sup_{n\to\infty} |H_n(x_n)| > 0$, lo cual contradice la derivabilidad de G.

A base de (6) y (8) obtenemos

$$\mathbf{P}(|H_n(w^n)| > \varepsilon) \leqslant \mathbf{P}(|H_n(w^n)| > \varepsilon, \quad w^n \in K_{h_n}) + \mathbf{P}(w^n \notin K_{h_n}),$$

$$\lim\sup_{n\to\infty} \mathbf{P}(|H_n(w^n)| > \varepsilon) \leqslant \delta.$$

Como δ es arbitrario, la relación (7) y junto con ella la afirmación del teorema quedan demostradas. ◁

Volvamos a examinar los ejemplos.

Sea η_p^* la cuantila muestral de orden p para la muestra Y de la distribución uniforme sobre [0, 1]. Entonces, de (5) y del teorema 1 obtenemos que

$$(\eta_p^* - p)\sqrt{n} \Rightarrow -w^\circ(p) \underset{d}{=} w^\circ(p).$$

Hemos determinado, además, que en el caso general, cuando F_0 es una función continua arbitraria de distribución, es válida la igualdad

$$\zeta_p^* = F_0^{-1}(\eta_p^*).$$

Si ahora utilizamos el tercer teorema de continuidad, obtendremos:

Corolario 1. *Si $X_n \in F_0$, F_0 es continuamente derivable en el punto ζ_p, $f(\zeta_p) = F_0'(\zeta_p) > 0$, entonces*

$$(\zeta_p^* - \zeta_p)\sqrt{n} \Rightarrow w^\circ(p)/f(\zeta_p).$$

Para la demostración sólo es necesario señalar que las condiciones del corolario 1 significan la derivabilidad continua de F_0^{-1} en el punto p,

$$(F_0^{-1}(p))' = \frac{1}{F_0'(F_0^{-1}(p))} = \frac{1}{f(\zeta_p)}.$$

Como $\mathbf{M}w^\circ(p) = 0$, $\mathbf{D}w^\circ(p) = \mathbf{M}(w(p) - pw(1))^2 = \mathbf{M}(w(p)(1 - p) + p(w(1) - w(p)))^2 = p(1 - p)^2 + p^2(1 - p) = p(1 - p)$, la afirmación del corolario 1 también puede escribirse en la forma

$$(\zeta_p^* - \zeta_p)\sqrt{n} \in \Phi_{0,\sigma^2}, \quad \sigma^2 = p(1 - p)/f^2(\zeta_p). \quad \triangleleft$$

En el ejemplo 2 derivamos la funcional $G(F) = \sup_{0\leqslant t\leqslant 1} |F(t) - F_0(t)|$ y, por lo tanto, según el teorema 1,

$$G(F_n^*)\sqrt{n} \Rightarrow \sup_{0\leqslant t\leqslant 1} |w^\circ(t)|.$$

Hemos hallado la distribución $\eta = \sup_{0 \le t \le 1} |w^\circ(t)|$ en forma explícita ([58]):

$$\mathbf{P}(\eta > z) = K(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2z^2}.$$

La función $K(z)$ se llama *función de Kolmogórov.*

Hemos visto que en el caso general, cuando F_0 es una función continua arbitraria de distribución, la distribución de la estadística

$$D(x) = \sup_t |F_n^*(t) - F_0(t)|$$

queda igual que para el caso $F_0(t) = t$, $t \in [0, 1]$. De este modo hemos obtenido:

Corolario 2 (teorema de Kolmogórov). *Si* $X \in F_0$, F_0 *es continua, entonces*

$$\sqrt{n}\, D(X) \in K.$$

Esto significa que la desviación máxima $D(X)$ de la función $F_n^*(t)$ de $F_0(t)$ tiene el orden $1/\sqrt{n}$ y puede representarse, aproximadamente, en la forma de $D(X) \approx \eta/\sqrt{n}$.

En el ejemplo 3 hemos visto que otra estadística (la cual a menudo se designa por ω^2)

$$\omega^2 = \int_{-\infty}^{\infty} (F_n^*(t) - F_0(t))^2 dF_0(t)$$

también es invariante respecto a F_0. Del teorema 1 se deduce:

Corolario 3. *Si* $X \in F_0$, F_0 *es continua, entonces*

$$n\omega^2 \Rightarrow \int_0^1 [w^\circ(t)]^2 dt.$$

La distribución $\int_0^1 [w^\circ(t)]^2 dt$ también fue hallada en forma explícita y, junto con la distribución $K(z)$, está tabulada. Con arreglo al ejemplo 4, el teorema 1 nos da:

Corolario 4. *Si* $X \in F_0$, F_0 *es continua, entonces*

$$\chi^2 \Rightarrow \sum_{j=1}^{r} (\delta_j w^\circ)^2/\delta_j,$$

donde δ_j, $j = 1, 2, \ldots, r$, *forman la partición del segmento* $[0, 1]$ *y están definidos en el ejemplo* 4.

Si suponemos que $\xi = (\xi_1, \ldots, \xi_r)$, $\xi_j = \delta_j w^\circ/\sqrt{\delta_j}$ utilizando el hecho

de que $\delta_j w^\circ = \delta_j w - w(1)\delta_j$, donde w es el proceso wieneriano estándar, obtenemos

$$\chi^2 \Rightarrow \sum_{j=1}^{r} \xi_j^2, \ \xi \in \Phi_{0,\sigma^2}.$$

Aquí $\sigma^2 = |\sigma_{ij}|$ es la misma matriz que en el ejemplo 7.3, puesto que

$$\delta_j w^\circ = \delta_j w - \left(\sum_k \delta_k w\right)\delta_j = \sum_{k=1}^{r} a_{kj}\delta_k w,$$

$$a_{kj} = \delta_{kj} - \delta_j, \ \mathbf{M}(\delta_k w)(\delta_l w) = \delta_{kl}\delta_k$$

(δ_{kl} es el símbolo de Kronecker),

$$\sigma_{ij} = \frac{\mathbf{M}(\delta_i w^\circ)(\delta_j w^\circ)}{\sqrt{\delta_i \delta_j}} = \frac{1}{\sqrt{\delta_i \delta_j}} \sum_{k=1}^{r} a_{ki} a_{kj} \delta_k =$$

$$= \frac{1}{\sqrt{\delta_i \delta_j}} (\delta_{ij}\delta_i - \delta_i \delta_j) = \delta_{ij} - \sqrt{\delta_i \delta_j}.$$

Repitiendo los razonamientos del ejemplo 7.3 obtenemos que $\sum_{j=1}^{r} \xi_j^2$ tiene una distribución χ^2 con $r-1$ grados de libertad.

Concluyendo este párrafo debemos señalar que no todas las estadísticas que representen interés pueden ser clasificadas como estadísticas de los tipos I ó II. Basta con examinar, por ejemplo, la estadística $S(X) = \sum_{i=1}^{n-1} x_i x_{i+1}$ o las estadísticas S relacionadas con las funcionales $G_n(F)$, donde las funcionales G_n dependen "considerablemente" de n (no sólo por la muestra), tales como, digamos, el término máximo de la serie variacional $S(X) = x_{(n)} = \zeta^*_{1-1/n}$ y otras.

§ 9*. Objeciones acerca de las estadísticas no paramétricas

Hay una propiedad respecto a la cual la estadística ζ_p^* en el ejemplo 8.1 se distingue considerablemente de las citadas en los ejemplos 8.2 — 8.4. Esta propiedad consiste en que la distribución límite de las estadísticas en los ejemplos 8.2 — 8.4 (véanse los corolarios 8.2 — 8.4) de ningún modo está relacionada con la función de distribución F_0, lo cual no se puede decir de la estadística ζ_p^* (compárese con el corolario 8.1).

Definición 1. La estadística $S(X)$ se llama *asintóticamente no paramétrica* si $S(X) \in Q$ cuando $n \to \infty$, y $\mathbf{Q}$ no depende de la distribución de X, o sea, no depende de F_0 si $X \in F_0$.

Cabe señalar que la propia función S en este caso puede depender de F_0. El término "no paramétrica" no es por sí mismo del todo acertado, no obstante, adquirió gran divulgación (está justificado en el caso en que F_0 pertenece a cierta familia paramétrica — entonces la distribución **Q** no depende del parámetro y desde este punto de vista no es paramétrica). A veces se utiliza otro término: "libre de la distribución".

En los §§ 6—8 hemos visto que las estadísticas $\sqrt{n}\,U(X)$, $\sqrt{n}\,D(X)$, $n\omega^2(X)$, $\chi^2(X)$ son asintóticamente no paramétricas.

También debemos indicar que el teorema 6.1 da la posibilidad de introducir un concepto más estrecho. En dicho teorema se ha establecido que $nF_n^*(t)$ está igual distribuida que $\eta(F_0(t))$, donde $\eta(u)$ es el proceso poissoniano convencional con un parámetro arbitrario $\lambda > 0$ a condición de que $\eta(1) = n$ (véase el § 6), o sea, dicho proceso no depende de F_0. Ahora bien, si la estadística S está construida como la funcional $G(F_n^*)$ (o $G(F_n^* - F_0)$), que es invariante respecto a la sustitución del "tiempo" t en el argumento, la distribución de S no dependerá de F_0. Por ejemplo,

$$D = \sup_t |F_n^*(t) - F_0(t)| = \frac{1}{n} \sup_t |\eta(F_0(t)) - nF_0(t)| =$$

$$= \frac{1}{n} \sup_{u \in [0,1]} |\eta(u) - un|. \qquad (1)$$

Lo dicho hace posible:

Definición 2. La estadística $S(X)$ se llama *no paramétrica* si su distribución no depende de $F_0 (X \in F_0)$.

Las relaciones (1) significan que la estadística D no es paramétrica.

También hemos señalado (véase el corolario 8.3) que la estadística ω^2, al igual que D, no depende de F_0 y, por lo tanto, tampoco es paramétrica.

La estadística χ^2, siendo asintóticamente no paramétrica no poseerá la propiedad de carácter no paramétrico. De esto es fácil convencerse directamente en un ejemplo, poniendo $r = 2$, $n = 1$.

Obtenemos otros ejemplos de las estadísticas no paramétricas si examinamos los valores de $F_n^*(\zeta_p)$, donde ζ_p es la cuantila de orden p, así que $nF_n^*(\zeta_p) \underset{d}{=} \eta(p)$ (véase el § 6). El número r_j de elementos de la muestra X, menores que x_j — la llamada estadística de rango — también será una estadística no paramétrica.

Los conceptos de estadísticas no paramétrica y asintóticamente no paramétrica son muy útiles en la teoría de la verificación de las hipótesis estadísticas (véase el capítulo 3), ya que la distribución de estas estadísticas, la cual es necesaria para la construcción de los criterios, es suficiente calcularla sólo una vez (por jemplo, para la distribución uniforme de F_0) y será útil para cualesquiera otras distribuciones de la muestra.

§ 10*. Distribuciones empíricas suavizadas. Densidades empíricas

En el § 2 a cada muestra X la hemos puesto en correspondencia con la distribución $\mathbf{P}_n^*$ que hemos llamado empírica y la cual no es más que la suma de n distribuciones atómicas concentradas en los puntos $x_1, \ldots, x_n$. Esta distribución posee varias propiedades magníficas descritas en los párrafos precedentes. Sin embargo, la definición de $\mathbf{P}_n^*$, utilizada por nosotros, no es la única posible ni mucho menos, y en varios casos no es la más natural. También existen otros puntos de vista en cuanto a la definición de $\mathbf{P}_n^*$, según los cuales las propiedades útiles (estudiadas anteriormente) de las distribuciones empíricas no sólo se conservan por completo, sino que son completadas por varias nuevas.

Aquí nos limitaremos a examinar la cuestión relacionada con la naturaleza de las distribuciones que situamos en los puntos x_i. En la definición de $\mathbf{P}_n^*$ que hemos utilizado, se trataba de las distribuciones degeneradas $\mathbf{I}_{x_i}(B)$, así que

$$\mathbf{P}_n^*(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{x_i}(B). \tag{1}$$

En este caso la distribución empírica es singular con respecto a la medida de Lebesgue y, por lo tanto, no tiene densidad. Esto puede resultar incómodo en los casos cuando sabemos de antemano que la distribución inicial $\mathbf{P}$ tiene densidad. Con arreglo a esta condición sería conveniente tener una distribución empírica suave $\mathbf{P}_n^*$ para la cual, junto con la convergencia $\mathbf{P}_n^* \to \mathbf{P}$, desde todos los puntos de vista establecidos anteriormente también tenga lugar la convergencia de las densidades $f_n^* \to f$, donde f_n^* y f son las densidades correspondientes a $\mathbf{P}_n^*$ y $\mathbf{P}$.

No es difícil obtener esto del modo siguiente. Sea $\mathbf{Q}$ cierta distribución *que tiene densidad.* Pongamos

$$\mathbf{P}_n^{**}(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Q}\left(\frac{B - x_i}{h_n}\right), \tag{2}$$

donde $\dfrac{B - x}{h}$ es el conjunto de puntos $y \in \mathscr{X}$ para los cuales $x + yh \in B$; $h_n \to 0$ cuando $n \to \infty$.

Es evidente que $\mathbf{P}_n^{**}(B)$ no es otra cosa sino la "suma media" de las distribuciones $\mathbf{Q}$ contraídas hasta las dimensiones h_n y "situadas" en los puntos x_i. La definición (2) generaliza (1). La fórmula (1) se obtiene de (2) si se pone $\mathbf{Q} = \mathbf{I}_0$, ya que $\mathbf{I}_{x_i}(B) = \mathbf{I}_0(B - x_i) = \mathbf{I}_0\left(\dfrac{B - x_i}{h_n}\right)$ para cualquier sucesión $\{h_n\}$.

Señalemos las siguientes propiedades de la distribución $\mathbf{P}_n^{**}$ que llamaremos *distribución empírica suavizada.*

1. La distribución $\mathbf{P}_n^{**}$ es la convolución de las distribuciones $\mathbf{P}_n^*$ y $\mathbf{Q}(B/h_n)$, y

$$\mathbf{P}_n(B) = \mathbf{MP}_n^{**}(B) = \int \mathbf{Q}\left(\frac{B-y}{h_n}\right)\mathbf{P}(dy)$$

es la convolución de las distribuciones $\mathbf{P}$ y $\mathbf{Q}(B/h_n)$. Con otras palabras, $\mathbf{P}_n(B)$ es la distribución de la variable aleatoria $\xi + h_n\eta$, donde $\xi \in \mathbf{P}$, $\eta \in \mathbf{Q}$. De los teoremas de continuidad se deduce que para $h_n \to 0$,

$$\mathbf{P}_n \Rightarrow \mathbf{P}. \tag{3}$$

Recordemos que para la distribución $\mathbf{P}_n^*$ hemos tenido la igualdad exacta

$$\mathbf{MP}_n^* = \mathbf{P}.$$

2. Si la distribución $\mathbf{P}$ es absolutamente continua con respecto a la medida de Lebesgue, la distribución $\mathbf{P}_n^{**}$ satisfará los teoremas análogos al de Glivenko — Cantelli. En efecto, en este caso la convergencia (3) significará la convergencia uniforme de las distribuciones sobre todos los intervalos. Para simplificar la exposición nos limitaremos a un caso unidimensional, supongamos que ($F_n^{**}(x)$, $F_n(x)$ y $Q(x)$ designan las funciones de distribución correspondientes a $\mathbf{P}_n^{**}$, $\mathbf{P}_n$ y $\mathbf{Q}$)

$$\begin{aligned} F_n^{**}(x) - F(x) &= \int \mathbf{Q}\left(\frac{x-y}{h_n}\right) dF_n^*(y) - F(x) = \\ &= -\int F_n^*(y) d_y Q\left(\frac{x-y}{h_n}\right) - F(x) = \\ &= F_n(x) - F(x) - \int (F_n^*(y) - F(y)) d_y Q\left(\frac{x-y}{h_n}\right). \end{aligned}$$

Aquí, como ya hemos señalado, la diferencia $F_n(x) - F(x) \to 0$ es uniforme en x, y la integral presente en el segundo miembro no excede $\sup_y |F_n^*(y) - F(y)| \underset{\text{c.s.}}{\to} 0$.

3. La ventaja de $\mathbf{P}_n^{**}$ en comparación con $\mathbf{P}_n^*$, por cuya razón hemos introducido la primera distribución, consiste en que *esta distribución tiene la densidad.*

$$f_n^*(x) = \frac{1}{nh_n}\sum_{i=1}^{n} q\left(\frac{x-x_i}{h_n}\right) = \frac{1}{h_n}\int q\left(\frac{x-y}{h_n}\right) dF_n^*(y) \tag{4}$$

($q(x)$ es la densidad de la distribución $\mathbf{Q}$) que para cada x, cuando $n \to \infty$ y $h_n \to 0$, se aproxima a la densidad $f(x)$ de la distribución $\mathbf{P}$.

Antes de demostrar la afirmación correspondiente, cabe señalar que para la obtención de buenos resultados acerca de la aproximación de $f_n^*(x)$ a $f(x)$, conviene utilizar las densidades limitadas suaves q. Al elegir, digamos, q indefinidas, la estimación $f_n^*(x)$ de la densidad suave $f(x)$ empeorará premeditadamente. Como la elección de q está en nuestras manos, podemos considerar que, por lo menos, queda cumplida la condición

$$d^2 = \int q^2(t)dt < \infty. \tag{5}$$

Teorema 1. *Si q satisface la condición* (5), *$f(x)$ es continua y limitada, $h_n \to 0$ para $n \to \infty$ de modo que $nh_n \to \infty$*, entonces

$$f_n^*(x) = f_n(x) + \zeta_n(x)/\sqrt{nh_n}, \tag{6}$$

donde $f_n(x)$ es la función no aleatoria

$$f_n(x) = \mathbf{M}f_n^*(x) = \mathbf{M}h_n^{-1}q\left(\frac{x - x_l}{h_n}\right) = \frac{1}{h_n}\int q\left(\frac{x-t}{h_n}\right) f(t)dt = \\ = \int q(z)f(x - zh_n)dz \to f(x) \tag{7}$$

para $h_n \to 0$. Las variables aleatorias $\zeta_n(x)$ son normales asintóticamente, $\zeta_n(x) \in \Phi_{0,\sigma^2(x)}$, $\sigma^2(x) = f(x)d^2$.

Demostración. La suma en (4) es la suma de variables aleatorias independientes e igualmente distribuidas en el esquema de series, con la particularidad de que $f_n(x) = \mathbf{M}f_n^*(x)$ está representada en (7). Pongamos

$$\xi_{k,n} = \frac{1}{\sqrt{nh_n}}\left[q\left(\frac{x - x_k}{h_n}\right) - h_n f_n(x)\right].$$

Entonces

$$f_n^*(x) - f_n(x) = \frac{1}{\sqrt{nh_n}}\sum_{k=1}^{n}\xi_{k,n}, \quad \mathbf{M}\xi_{k,n} = 0$$

$$\mathbf{M}\xi_{k,n}^2 = \frac{1}{n}\left[\mathbf{M}\frac{1}{h_n}q^2\left(\frac{x - x_k}{h_n}\right) - h_n f_n^2(x)\right],$$

$$\mathbf{M}\frac{1}{h_n}q^2\left(\frac{x - x_k}{h_n}\right) = \frac{1}{h_n}\int q^2\left(\frac{x-t}{h_n}\right)f(t)dt = \\ = \int q^2(z)f(x - zh_n)dz \to f(x)\int q^2(z)dz = f(x)d^2. \tag{8}$$

Ahora bien, $\mathbf{M}\xi_{k,n}^2 \sim f(x)d^2/n$ si $f(x) > 0$. La condición de Lindeberg tiene en nuestro caso la forma

$$n\mathbf{M}(\xi_{1,n}^2;\ |\xi_{1,n}| > \varepsilon) \to 0 \tag{9}$$

para $n \to \infty$ y para cualquier $\varepsilon > 0$. Como $h_n f_n^2(x) \to 0$, $n\xi_{1,n}^2 \leqslant 2(q^2((x - x_1)/h_n) + h_n f_n^2(x))$, entonces para cumplir (9) es suficiente que

$$\mathbf{M}\left(\frac{1}{h_n} q^2\left(\frac{x - x_1}{h_n}\right)\right); \; q\left(\frac{x - x_1}{h_n}\right) > \varepsilon\sqrt{nh_n} \to 0.$$

Esta relación tiene lugar, ya que su primer miembro es igual a (compárese con (8))

$$\int\limits_{q(z) > \varepsilon\sqrt{nh_n}} q^2(z) f(x - zh_n) dz \leqslant c \int\limits_{q(z) > \varepsilon\sqrt{nh_n}} q^2(z) dz \to 0.$$

Ahora bien, a la variable aleatoria $\zeta_n(x) = \sum\limits_{k=1}^{n} \xi_{k,n}$ es aplicable el teorema central del límite. Esto demuestra el teorema 1. ◁

En el problema sujeto a examen surge naturalmente la cuestión acerca de la elección óptima de h_n y de la función $q(t)$. Sin embargo, su solución depende de las propiedades de suavidad de $f(x)$. En efecto, supongamos, por ejemplo, que $f(x)$ es positiva solamente en el intervalo finito y que es dos veces continuamente derivable con el valor fijo $\varphi = \int (f''(x))^2 dx$. Supongamos también que $\int zq(z)dz = 0$ (esto es siempre así para las $q(z)$ simétricas) y que $D^2 = \int z^2 q(z)dz < \infty$. Entonces

$$f_n(x) = \int q(z) f(x - zh_n) dz =$$

$$= \int q(z) \left[f(x) - zh_n f'(x) + \frac{z^2 h_n^2}{2} f''(x) + o(z^2 h_n^2)\right] dz =$$

$$= f(x) + \frac{h_n^2 f''(x)}{2} \int z^2 q(z) dz + o(h_n^2).$$

Vemos que

$$f_n^*(x) - f(x) = \frac{D^2 h_n^2 f''(x)}{2} + \frac{\zeta_n(x)}{\sqrt{nh_n}} + o(h_n^2),$$

$$\mathbf{M}[f_n^*(x) - f(x)]^2 = \left(\frac{D^2 h_n^2 f''(x)}{2}\right)^2 + \frac{d^2 f(x)}{nh_n} + o(h_n^4). \qquad (10)$$

La minimización de esta expresión en h_n y q dará, en virtud de la normalidad asintótica de $\zeta_n(x)$, la "dispersión" mínima posible de $f_n^*(x)$ alrededor del valor de $f(x)$. No obstante, en este caso los valores minimizantes de h_n y q dependerán de x mediante los valores desconocidos de $f(x)$ y $f''(x)$. Para evitar este efecto y obtener la optimalidad "por término medio" es

natural examinar la integral

$$\int \mathbf{M}[f_n^*(x) - f(x)]^2 dx \tag{11}$$

cuya parte principal será igual a $\left(\frac{D^2 h_n^2}{2}\right)^2 \varphi + \frac{d^2}{nh_n}$ (esto se obtiene si en (10) se retira $o(h_n^4)$).

El mínimo de esta expresión se alcanza cuando $h_n = \left(\frac{d^2}{nD^4\varphi}\right)^{1/5}$. Con tal elección de h_n, la integral (11) será igual a

$$\frac{5}{4}\varphi^{1/5}(Dd^2)^{4/5}n^{-4/5} + o(n^{-4/5}), \tag{12}$$

$$-f_n^*(x) - f(x) = \left(\frac{Dd^2}{n\varphi}\right)^{2/5}\left(\frac{f''(x)}{2} + f(x)\sqrt{\varphi}\xi_n + o(n^{-2/5})\right),$$

$$\xi_n \in \Phi_{0,1}.$$

Ahora bien, aquí la velocidad de convergencia constituye sólo $n^{-2/5}$ a diferencia de la velocidad $n^{-1/2}$, la cual tiene lugar para la convergencia de las funciones de distribución. Es un hecho natural, ya que en la estimación del valor de $f(x)$ toma parte, hablando en términos generales, no toda la muestra, sino las observaciones que se han concentrado en cierto entorno decreciente del punto x.

La expresión (12) permite también elegir del modo óptimo la función $q(z)$, o sea, la función para la cual se minimiza Dd^2. Suponiendo, sin limitar la generalidad, que $D = 1$, obtenemos el problema de minimización $d^2 = \int q^2(z)dz$ a condición de que $\int q(z)dz = \int z^2 q(z)dz = 1$, $\int zq(z)dz = 0$.

Nótese que si f tiene derivadas continuas de orden más alto que $2m > 2$, también pueden obtenerse velocidades más altas de convergencia de la diferencia $f_n^*(x) - f(x)$ hacia cero. Sin embargo, en este caso es necesario utilizar las distribuciones generalizadas $\mathbf{Q}$ cuya "densidad" q puede tomar los valores de ambos signos y permite satisfacer las condiciones $\int z^{2m} q(z)dz = 1$, $\int z^j q(z)dz = 0$ para todos los $1 \leqslant j \leqslant 2m - 1$. En este caso, mediante los razonamientos anteriores podemos obtener la velocidad de convergencia de orden de $n^{-\frac{2m}{4m+1}} = n^{-1/2 + \frac{1}{2(4m+1)}}$ la cual será tanto mejor cuanto mayor sea m. Este hecho se explica por la circunstancia de que para $f(x)$ más suaves, en la estimación del valor de $f(x)$ se incorporan los elementos de la muestra, situados en entornos cada vez más amplios del punto x.

Por otro lado, eligiendo funciones suaves $q(z)$, podemos asegurar la posibilidad de estimar no sólo las densidades $f(x)$, sino también sus derivadas. De esto también podemos convencernos a base de los razonamientos anteriormente citados.

La función $f_n^*(x)$, que tiene la forma (4), se llama frecuentemente estimación de Rosenblatt — Parzen de la densidad $f(x)$ o *estimación nuclear de* $f(x)$. En este caso las funciones $q(z)$ se llaman *núcleos*. En la práctica se utilizan a menudo los núcleos "rectangulares", o sea, se supone que

$$q(z) = \begin{cases} 1 \text{ para } z \in [-1/2,\ 1/2], \\ 0 \text{ para } z \notin [-1/2,\ 1/2]. \end{cases}$$

A veces se procede de un modo todavía más sencillo: la recta real se divide en pequeños intervalos Δ_j (de h_n de largo) y se supone que $f_n^*(x) = \dfrac{\nu_j}{nh_n}$ para $x \in \Delta_j$, donde ν_j es el número de elementos de la muestra que coincidieron con Δ_j. Tal función $f_n^*(x)$ se llama *histograma* de la muestra. Es fácil comprobar que si $f(x)$ es continua, entonces el histograma $f_n^*(x)$, a la par con la función definida en (4), también posee la propiedad de convergencia $f_n^*(x) \xrightarrow[p]{} f(x)$ si $h_n \to 0$, $nh_n \to \infty$.

CAPÍTULO 2

Teoría de estimación de los parámetros desconocidos

El § 2 contiene la descripción de las familias paramétricas más difundidas de distribuciones y sus propiedades principales.

En los §§ 3—6 se exponen métodos principales de obtención de las estimaciones puntuales.

En los §§ 7 y 8 se examinan los enfoques de la comparación de las estimaciones.

Los §§ 9—20 están dedicados a los métodos de construcción de las estimaciones óptimas (en uno u otro sentido). Se destacan las cuatro direcciones siguientes:

1) (§§ 9—11 y 20) Enfoques bayesiano y minimax de la construcción de las estimaciones. Los §§ 9 y 10 son de carácter adicional y contienen las definiciones y la exposición de las propiedades principales de las esperanzas matemáticas condicionales y de las distribuciones condicionales.

2) (§§ 12—15) Construcción de las estimaciones óptimas (eficientes) con ayuda de los principios de suficiencia y de no desplazamiento.

3) (§§ 16, 17 y 22) Construcción de las estimaciones óptimas (eficientes) basándose en la desigualdad de Rao — Cramer.

4) (§§ 18 y 19) Utilización de las consideraciones de invariación.

En los §§ 21—29 se estudian las propiedades asintóticas de la relación de verosimilitud. Sobre esta base se determina la optimación asintótica de las estimaciones de verosimilitud. Los resultados de los §§ 21—29 también constituyen la base de la teoría de los criterios óptimos, desarrollada en el capítulo 3.

Los §§ 31 y 32 están dedicados a la estimación por intervalos.

§ 1. Observaciones preliminares

Como ya hemos señalado en los párrafos precedentes, el objeto inicial de las investigaciones estadísticas está constituido por la muestra

$$X_n = (x_1, \ldots, x_n),\ x_i \in \mathscr{X},$$

de la distribución $\mathbf{P}$, la cual es desconocida por completo o parcialmente. En la estadística matemática se destacan, en calidad de principales, las dos siguientes clases de problemas:

1. *Estimación de los parámetros desconocidos.*

2. *Verificación de las hipótesis estadísticas.*

Los problemas de primera clase aparecen cuando por la muestra $X = X_n$ es necesario estimar cualquier característica numérica desconocida θ de la distribución $\mathbf{P}$ (que ya es desconocida). O sea, para la funcional dada

$$\theta = \theta(\mathbf{P}),$$

de la distribución $\mathbf{P}$ debemos señalar la función de la muestra (o bien, que es lo mismo, la estadística)

$$\theta^* = \theta_n^*(X_n)$$

destinada a la utilización, en vez del parámetro θ, en calidad de su aproximación. En el capítulo precedente hemos visto que las premisas para esto existen. La estadística θ^* se llama *estimación* del parámetro θ. Claro está que las estimaciones para el parámetro θ pueden ser muchísimas. El teorema 1.3.1 muestra que, por ejemplo, para la estimación de la funcional $\theta = \theta(\mathbf{P})$, que tiene la forma

$$\theta = \int g(x)\, dF(x),$$

es natural utilizar la estadística

$$\theta^* = \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Pero claro que también se pueden examinar otras estimaciones, digamos,

$$\theta^* = \frac{1}{n - \nu_1 - \nu_2} \sum_{j=\nu_1+1}^{n-\nu_2} g(x_{(j)}),$$

donde $x_{(j)}$, $j = 1, \ldots, n$, son los elementos de la serie variacional, etc. En calidad de θ^* también pueden tomarse los valores que no dependen de la muestra. Se puede poner, por ejemplo, $\theta^* \equiv 0$, aunque esto no siempre es racional y es completamente irracional cuando el conjunto de valores posibles de θ no contiene 0.

En relación con la última observación es preciso señalar que en el planteamiento del problema sobre la estimación se indica con frecuencia cuál es el conjunto Θ de los valores posibles de θ. Por ejemplo, si se aprecia la porción θ de un mineral cualquiera contenido en la mena, entonces, claro está que $\theta \in [0, 1]$.

En muchos casos también se sabe de antemano que la distribución $\mathbf{P}$ de la muestra X no puede ser arbitraria, sino que pertenece a una familia determinada de distribuciones $\mathscr{P}$.

Entre los problemas de la estimación de los parámetros figura el ejemplo 1 dado en la Introducción.

Los problemas de segunda clase se refieren a la comprobación de una y otra suposición (hipótesis) sobre la distribución desconocida **P**. Por ejemplo, podemos verificar la hipótesis consistente en que **P** tiene una u otra forma dada. A este tipo de problemas pertenece el ejemplo 2 dado en la Introducción.

Más tarde veremos que no hay diferencia cualitativa entre los problemas de primera clase (teoría de las estimaciones) y de segunda clase (verificación de las hipótesis estadísticas).

En este capítulo expondremos los planteamientos de los problemas y los enfoques que están íntimamente vinculados con los resultados del capítulo precedente y que pueden llamarse "puramente estadísticos" a distinción de los enfoques más generales de la teoría de los juegos, que se examinan en el cap. 5.

Los enfoques puramente estadísticos expresan, en cierta medida, la esencia de los métodos de la estadística matemática. Históricamente tales enfoques fueron comprendidos mucho antes que los métodos más generales. En cuando a su aplicación, por lo visto, el hombre los utilizaba explícita o implícitamente a lo largo de todo el proceso del conocimiento.

Todo esto justifica la exposición independiente de los enfoques puramente estadísticos, a pesar de que ciertos momentos de esta exposición pueden considerarse como casos particulares en el marco de las concepciones más generales. Al mismo tiempo revelaremos cierta insuficiencia del enfoque puramente estadístico para planteamientos más exactos de los problemas. Esto nos ayudará a comprender el carácter racional de otros puntos de vista.

§ 2. Algunas familias paramétricas de distribuciones y sus propiedades

Examinemos algunas familias de distribuciones que dependen de los parámetros (o familias paramétricas de distribuciones) que con frecuencia surgen en los suplementos y que aparecerán en la exposición ulterior tanto de hecho como en calidad de ilustraciones.

1. Distribución normal en una recta. Con el símbolo Φ_{α,σ^2} designaremos la distribución normal con los parámetros (α, σ^2), o sea, la distribución de densidad

$$\varphi_{\alpha,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\alpha)^2}{2\sigma^2}},$$

así que

$$\Phi_{\alpha,\sigma^2}(B) = \int_B \varphi_{\alpha,\sigma^2}(x)dx.$$

Si $\xi \in \Phi_{0,1}$ y $k \geqslant 0$ es un número entero, entonces, evidentemente,

$$\mathbf{M}\xi^{2k+1} = 0.$$

Para los momentos de orden par, utilizando la sustitución $x = \sqrt{2}u$, encontramos

$$\mathbf{M}\xi^{2k} = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k} e^{-x^2/2} dx = \frac{2^{k+1}}{\sqrt{2\pi}} \int_0^\infty u^k e^{-u} \frac{du}{\sqrt{2u}} =$$

$$= \frac{2^k}{\sqrt{\pi}} \Gamma(k + 1/2), \qquad (1)$$

donde $\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx$ es la función Γ, $\Gamma(\lambda) = (\lambda - 1)\Gamma(\lambda - 1)$, $\Gamma(1/2) = \sqrt{\pi}$, así que

$$\mathbf{M}\xi^{2k} = (2k - 1)!! = (2k - 1)(2k - 3) \ldots 1.$$

También obtendríamos este resultado si hubiéramos derivado $2k$ veces la función característica $e^{-t^2/2}$ en el punto $t = 0$.

2. Distribución normal multidimensional. En el caso multidimensional $\mathscr{X} = R^m$, el símbolo Φ_{α,σ^2} significará la distribución normal en R^m con el vector de esperanzas matemáticas $\alpha = (\alpha_1, \ldots, \alpha_m)$ y con la matriz de segundos momentos centrales $\sigma^2 = \|\sigma_{ij}\|$, $i, j = 1, \ldots, m$. Si A es la matriz inversa a σ^2 (en los casos cuando ella existe), entonces la densidad $\varphi_{\alpha,\sigma^2}(x)$ en R^m de la distribución Φ_{α,σ^2} tiene la forma a (véase [11], p. 148)

$$\varphi_{\alpha,\sigma^2}(x) = \frac{\sqrt{|A|}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}(x - \alpha)A(x - \alpha)^T\right),$$

donde x^T es el vector transpuesto. Recordemos también (ya hemos utilizado este hecho en el § 1.7) que la función característica de la magnitud $\xi \in \Phi_{\alpha,\sigma^2}$ es igual a

$$\mathbf{M}e^{it\xi^T} = \exp\left(it\alpha^T - \frac{1}{2} t\sigma^2 t^T\right),$$

donde $t = (t_1, \ldots, t_m)$ es el vector en R^m.

3. Distribución gamma. El símbolo $\Gamma_{\alpha,\lambda}$ designará la llamada "distribución gamma" (o distribución Γ) con los parámetros (α, λ). La densidad $\gamma_{\alpha,\lambda}(x)$ de esta distribución depende de dos parámetros $\alpha > 0$ y $\lambda > 0$ y es igual a (véase [11] y § 7 del cap. 6)

$$\gamma_{\alpha,\lambda}(x) = \begin{cases} \dfrac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, & x \geqslant 0, \\ 0 & x < 0, \end{cases} \qquad (2)$$

donde $\Gamma(\lambda)$ es la función Γ definida en (1). La función característica de la distribución Γ tiene la forma ([11])

$$\int_0^\infty e^{itx}\gamma_{\alpha,\lambda}(x)dx = \left(1 - \frac{it}{\alpha}\right)^{-\lambda}. \tag{3}$$

Si $\xi \in \Gamma_{\alpha,\lambda}$, entonces

$$\mathbf{M}\xi^t = \frac{\alpha^\lambda}{\Gamma(\lambda)}\int_0^\infty x^{\lambda+t-1}e^{-\alpha x}dx = \frac{\alpha^{-t}}{\Gamma(\lambda)}\int_0^\infty y^{\lambda+t-1}e^{-y}dy = \frac{\alpha^{-t}\Gamma(\lambda+t)}{\Gamma(\lambda)}. \tag{4}$$

Para enteros $t > 0$, el mismo resultado podría ser obtenido derivando la función característica. Poniendo $t = 1, 2$, encontramos

$$\mathbf{M}\xi = \lambda/\alpha, \quad \mathbf{D}\xi = \lambda/\alpha^2. \tag{5}$$

De las fórmulas (3) y (4) se deduce que el parámetro α desempeña el papel de escala, así que

$$\eta/\alpha \in \Gamma_{\alpha,\lambda} \quad \text{si} \quad \eta \in \Gamma_{1,\lambda}.$$

En virtud de esta circunstancia, muchas propiedades de la distribución Γ pueden ser estudiadas para un valor cualquiera de α, por ejemplo, para $\alpha = 1$ o para $\alpha = 1/2$. A menudo el segundo valor será para nosotros más cómodo, ya que la distribución $\Gamma_{1/2,\lambda}$ desempeña un importante papel independiente en la estadística matemática y se llama distribución "ji-cuadrado" (o distribución χ^2).

4. Distribución "ji-cuadrado" $\mathbf{H}_k$ con k grados de libertad. Así se denomina la distribución $\mathbf{H}_k = \Gamma_{1/2,\,k/2}$ cuando $k > 0$ son enteros. Conservaremos esta denominación para la distribución $\mathbf{H}_k$ cuando también se trate de $k > 0$ arbitrarios. En virtud de (3), la función característica de la distribución $\mathbf{H}_k$ es igual a

$$(1 - 2it)^{-k/2}.$$

Indiquemos las tres siguientes propiedades de la distribución $\mathbf{H}_k$.

1) *Si η_i son independientes, $\eta_i \in \mathbf{H}_{k_i}$, $i = 1, \ldots, s$, entonces*

$$\sum_{i=1}^{s}\eta_i \in \mathbf{H}_k, \quad k = \sum_{i=1}^{s} k_i.$$

Esta propiedad se deduce directamente de la forma de la función característica de la distribución $\mathbf{H}_k$.

2) *Si $\xi \in \Phi_{\alpha,\sigma^2}$, donde Φ_{α,σ^2} es la distribución normal k-dimensional con la matriz no degenerada de segundos momentos σ^2, entonces*

$$Q(\xi) = (\xi - \alpha)\sigma^{-2}(\xi - \alpha)^T \in \mathbf{H}_k.$$

En efecto, la función característica de la variable aleatoria $Q(\xi)$ es igual a

$$\mathbf{M}e^{itQ(\xi)} = \frac{\sqrt{|\sigma^{-2}|}}{(2\pi)^{k/2}} \int \exp\left(-\frac{1}{2} Q(x)(1 - 2it)\right) dx_1, \ \ldots \ dx_k.$$

Sustituyendo las variables $x_j \sqrt{1 - 2it} = y_j$, obtenemos la expresión

$$(1 - 2it)^{-k/2} \frac{\sqrt{|\sigma^{-2}|}}{(2\pi)^{k/2}} \int e^{-\frac{1}{2} Q(y)} dy_1, \ \ldots \ dy_k = (1 - 2it)^{-k/2},$$

que es lo que se necesitaba demostrar. El hecho de que la integral en el primer miembro no depende de la variación del dominio de integración se deriva de la analiticidad de la función subintegral y de su decrecimiento rápido cuando $|y| \to \infty$ (compárese con [11], p. 131).

De lo dicho resulta que la distribución $\mathbf{H}_k$ está contenida en la variable aleatoria

$$\chi^2 = \xi_1^2 + \ldots + \xi_k^2,$$

donde ξ_j son independientes, $\xi_j \in \Phi_{0,1}$. El término "número de grados de libertad" está precisamente relacionado con esta representación.

3) Como $\mathbf{M}\xi_1^2 = 1$, $\mathbf{M}\xi_1^4 = 3$, $D\xi_1^2 = 2$ para $\xi_1 \in \Phi_{0,1}$, *entonces, en virtud del teorema central del límite, para* $k \to \infty$,

$$\frac{\chi^2 - k}{\sqrt{2k}} \in \Phi_{0,1}. \tag{6}$$

De aquí y de los teoremas de continuidad enunciados en el § 1.5 se deduce que a la par con (6),

$$\sqrt{2\chi^2} - \sqrt{2k - 1} \in \Phi_{0,1}.$$

Esta convergencia sirve de base para la igualdad aproximada (en caso de k y x grandes) $\mathbf{H}_k(0, x) \approx \Phi(\sqrt{2x} - \sqrt{2k - 1})$, $\Phi(x) = \Phi_{0,1}((-\infty, x))$, la cual, por regla general, resulta más exacta que la aproximación $\mathbf{H}_k((0, x) \approx \Phi\left(\frac{x - k}{\sqrt{2k}}\right)$ que se deduce de (6).

Señalemos otro caso particular de la distribución Γ, el cual aparece a menudo en las aplicaciones.

5. Distribución exponencial. Es la distribución $\Gamma_{\alpha,1}$ de densidad

$$\alpha e^{-\alpha x}, \ x > 0.$$

De las fórmulas (5) obtenemos, para $\xi \in \Gamma_{\alpha,1}$,

$$\mathbf{M}\xi = 1/\alpha, \quad D\xi = 1/\alpha^2.$$

Examinemos ahora ciertas distribuciones relacionadas con las distribuciones normal y gamma y que desempeñan un papel importante en la estadística matemática. A distinción de las anteriores, con estas distribuciones no hemos tropezado anteriormente.

6. Distribución de Fisher $\mathbf{F}_{k_1 k_2}$ con k_1 y k_2 número de grados de libertad. Así se llama la distribución de la variable aleatoria

$$\zeta = \eta_1/\eta_2,$$

donde η_j son independientes, $\eta_j \in \mathbf{H}_{k_j}$, $j = 1, 2$. De las propiedades de la distribución Γ se deduce que la distribución de ζ queda igual cuando $\eta_j \in \Gamma_{\alpha,\, k_j/2}$ y para cualquier $\alpha > 0$, y que ζ cuando k_j son enteros, admite la representación

$$\zeta = \frac{\xi_1^2 + \ldots + \xi_{k_1}^2}{\zeta_1^2 + \ldots + \zeta_{k_2}^2},$$

donde las variables aleatorias ξ_i, ζ_k son independientes, $\xi_i \in \Phi_{0,1}$, $\zeta_k \in \Phi_{0,1}$.

Hallemos la densidad de la distribución $\mathbf{F}_{k_1,k_2}$. Tenemos

$$\mathbf{P}(\zeta < x) = \iint\limits_{u/v < x} \Gamma_{1,\lambda_1}(du)\Gamma_{1,\lambda_2}(dv) = \int\limits_{v=0}^{\infty} \int\limits_{u=0}^{vx} \frac{u^{\lambda_1 - 1} v^{\lambda_2 - 1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)} e^{-u-v} du\, dv;$$

$$f_{(\zeta)}(x) = \frac{d\mathbf{P}(\zeta < x)}{dx} = \int\limits_0^{\infty} \frac{(vx)^{\lambda_1 - 1} v^{\lambda_2 - 1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)} e^{-v - vx} v\, dv =$$

$$= \frac{x^{\lambda_1 - 1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \int\limits_0^{\infty} v^{\lambda_1 + \lambda_2 - 1} e^{-v(1+x)} dv = \frac{x^{\lambda_1 - 1}\Gamma(\lambda_1 + \lambda_2)}{(1+x)^{\lambda_1 + \lambda_2}\Gamma(\lambda_1)\Gamma(\lambda_2)}. \quad (7)$$

Es evidente que la densidad necesaria se obtiene si aquí se sustituye $\lambda_j = k_j/2$. Es fácil determinar los momentos de la variable aleatoria ζ (si éstos existen):

$$\mathbf{M}\zeta^l = \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \int\limits_0^{\infty} \frac{x^{\lambda_1 + l - 1}}{(1+x)^{\lambda_1 + \lambda_2}} dx = \frac{\Gamma(\lambda_1 + l)\Gamma(\lambda_2 - l)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}. \quad (8)$$

En particular, cuando $l = 1, 2$, obtenemos

$$\mathbf{M}\zeta = \frac{\lambda_1}{\lambda_2 - 1}, \quad \mathbf{M}\zeta^2 = \frac{\lambda_1(\lambda_1 + 1)}{(\lambda_2 - 1)(\lambda_2 - 2)}.$$

La distribución de Fisher también a veces se llama distribución de Snedecor. Esto se debe al hecho de que Fisher propuso utilizar y tabuló, en

realidad, no la distribución de ζ, sino la de la variable aleatoria $\frac{1}{2}\ln\zeta$. En cuanto a la distribución de ζ, ésta fue tabulada un poco más tarde por Snedecor

7. Distribución de Student *)T_k con k grados de libertad. Esta es, por definición, la distribución de la variable aleatoria

$$t = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2)}},$$

donde ξ_j son independientes, $\xi_j \in \Phi_{0,1}$, $j = 0, \ldots, k$. Es evidente que $-t$ tiene la misma distribución y, por lo tanto, la distribución de Student es simétrica con respecto al origen de coordenadas. Luego

$$t^2 = \frac{k\xi_0^2}{\xi_1^2 + \ldots + \xi_k^2} = \frac{k\eta_1}{\eta_2},$$

donde η_j son independientes, $\eta_1 \in \mathbf{H}_1$, $\eta_2 \in \mathbf{H}_k$. Esto quiere decir que t^2/k tiene la distribución de Fisher. Examinemos la variable aleatoria $\tau = \sqrt{\zeta}$, $\zeta = \eta_1/\eta_2$, $\eta_j \in \mathbf{H}_{k_j}$. Como $\mathbf{P}(\tau < x) = \mathbf{P}(\zeta < x^2)$, la densidad $f_{(\tau)}(x)$ de la variable aleatoria τ será igual a

$$f_{(\tau)}(x) = 2x f_{(\zeta)}(x^2) = 2x\,\frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \cdot \frac{x^{2\lambda_1-2}}{(1+x^2)^{\lambda_1+\lambda_2}} =$$

$$= \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \cdot \frac{2x^{2\lambda_1-1}}{(1+x^2)^{\lambda_1+\lambda_2}}, \quad \lambda_j = k_j/2, \quad x > 0.$$

De aquí, cuando $\lambda_1 = 1/2$, $\lambda_2 = k/2$, se puede obtener, de un modo evidente, la densidad $|t|/\sqrt{k}$. Como la distribución de t es simétrica, para la densidad $f_{(t)}(x)$ de la variable aleatoria t tenemos finalmente

$$f_{(t)}(x) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\,\Gamma(k/2)}\left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}. \tag{9}$$

Por supuesto que todos los momentos de t de orden impar (si existen) son iguales a cero. Para los momentos de orden par $2l$ tenemos, en virtud de (8),

$$\mathbf{M}t^{2l} = k^l \mathbf{M}\zeta^l = k^l\,\frac{\Gamma(\lambda_1+l)\Gamma(\lambda_2-l)}{\Gamma(\lambda_1)\Gamma(\lambda_2)},$$

donde es necesario poner $\lambda_1 = 1/2$, $\lambda_2 = k/2$, $2l < k$. Si $l = 1$ obtenemos

$$\mathbf{M}t^2 = \frac{k}{k-2}.$$

*) *Student* es el seudónimo de W. S. Gosset.

Según su forma, la función $f_{(t)}(x)$ se parece a la densidad de la ley normal. Además, con el crecimiento de k,

$$f_{(t)}(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

que significa la convergencia $t \in \Phi_{0,1}$ cuando $k \to \infty$. Sin embargo, $f_{(t)}(x)$ tiene "colas más gruesas", puesto que con el aumento de $|x|$, la función (9) disminuye mucho más lentamente que $e^{-x^2/2}$, así que para todos $b > 0$,

$$\mathbf{T}_k((-b, b)) < \Phi_{0,1}((-b, b)). \tag{10}$$

En este caso, la diferencia entre el segundo y el primer miembro en (10) puede ser considerable cuando k no son grandes.

El lector también puede demostrar la convergencia $t = \sqrt{k}\xi_0 / \sqrt{\eta_2}$ hacia la ley normal, utilizando otra vía, por medio del teorema de continuidad. Por ejemplo, basta con notar que $\frac{\eta_2}{k} = \frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2) \xrightarrow[c.s.]{} 1$ y, por lo tanto, $t \xrightarrow[c.s.]{} \xi_0$, $t \Rightarrow \xi_0$.

8. Distribución beta (B-distribución). Así se llama la distribución $\mathbf{B}\lambda_1, \lambda_2$ de densidad

$$f_{(\beta)}(x) = \begin{cases} \dfrac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} x^{\lambda_1 - 1}(1 - x)^{\lambda_2 < 1}, & x \in [0, 1], \\ 0, & x \notin [0, 1] \end{cases}$$

Se denomina así debido a la función beta

$$\mathrm{B}(\lambda_1, \lambda_2) = \int_0^1 x^{\lambda_1 - 1}(1 - x)^{\lambda_2 - 1} dx = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1 + \lambda_2)}.$$

La distribución beta está relacionada con la distribución gamma y la distribución de Fisher por medio de la afirmación siguiente:

Si η_j son independientes, $\eta_j \in \Gamma_{\alpha, \lambda_j}$ (o bien $\eta_j \in \mathbf{H}_{2\lambda_j}$), entonces

$$\beta = \frac{\eta_1}{\eta_1 + \eta_2} = \frac{\zeta}{\zeta + 1} \in \mathbf{B}_{\lambda_1, \lambda_2},$$

donde $\zeta = \eta_1 / \eta_2 \in \mathbf{F}_{2\lambda_1, 2\lambda_2}$.

La demostración de esta afirmación es muy fácil, ya que en virtud de (7), $\mathbf{P}(\beta < x) = \mathbf{P}\left(\zeta < \dfrac{x}{1 - x}\right)$,

$$f_{(\beta)}(x) = f_{(\beta)}\left(\frac{x}{1-x}\right)\left(\frac{x}{1-x}\right)' = \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\left(\frac{x}{1-x}\right)^{\lambda_1-1} \times$$

$$\times(1-x)^{\lambda_1+\lambda_2-2} = \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, x^{\lambda_1-1}(1-x)^{\lambda_2-1},\ x \in [0,\ 1].$$

Para los momentos de la variable aleatoria β tenemos

$$\mathbf{M}\beta^l = \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\int_0^1 x^{\lambda_1+l-1}(1-x)^{\lambda_2-1}dx = \frac{\Gamma(\lambda_1+\lambda_2)\Gamma(\lambda_1+l)}{\Gamma(\lambda_1)\Gamma(\lambda_1+\lambda_2+l)}\,.$$

Para $l = 1,\ 2$ obtenemos

$$\mathbf{M}\beta = \frac{\lambda_1}{\lambda_1+\lambda_2}\,,\quad \mathbf{M}\beta^2 = \frac{\lambda_1(\lambda_1+1)}{(\lambda_1+\lambda_2)(\lambda_1+\lambda_2+1)}\,.$$

9. Distribución uniforme. La distribución uniforme sobre [0, 1], que se obtiene si se pone $\lambda_1 = \lambda_2 = 1$, es un caso particular de la B-distribución.

Designaremos con el símbolo $\mathbf{U}_{a,\,b}$ la distribución uniforme sobre el segmento $[a,\ b]$, así que $\mathbf{B}_{1,1} = \mathbf{U}_{0,1}$.

Con ayuda de B-distribución se puede describir la distribución de los términos de la serie variacional $x_{(k)}$ de la muestra X.

Teorema 1. *Si $X \in \mathbf{P}$ es la muestra de la distribución $\mathbf{P}$ con la función continua de distribución F, entonces*

$$y_{(k)} = F(x_{(k)}) \in \mathbf{B}_{k,\,n-k+1}.$$

Demostración. Como $y_k = F(x_k) \in \mathbf{U}_{0,\,1}$, entonces $y_{(k)} = F(x_{(k)})$ puede considerarse como término de la serie variacional de la muestra $Y \in \mathbf{U}_{0,1}$. Determinemos $\mathbf{P}(y_{(k)} \in (u,u+du))$. El suceso $\{y_{(k)} \in (u,\ u+du)$ se puede representar como la unión de los sucesos disjuntos

$$A_j = \{y_j \in (u,\ u+du),\ y_j = y_{(k)}\},$$

que se producen cuando y_j adquiere el valor de $(u,\ u+du)$ (esta probabilidad es igual a du), cuando $k-1$ observaciones, de las $n-1$ restantes, caen en el campo de valores de $(0,\ u)$, y cuando $n-k$ observaciones caen en el campo de valores de $(u,\ 1)$. Por consiguiente,

$$\mathbf{P}(A_j) = C_{n-1}^{k-1}u^{k-1}(1-u)^{n-k}du,$$

$$\mathbf{P}(y_{(k)} \in (u,\ u+du)) = nC_{n-1}^{k-1}u^{k-1}(1-u)^{n-k}du.$$

Esto precisamente significa que la densidad $y_{(k)}$ existe y es igual a

$$\frac{n!}{(k-1)!(n-k)!}\,u^{k-1}(1-u)^{n-k} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\,u^{k-1}(1-u)^{n-k}.\triangleleft$$

Basándose en el teorema 1 también es fácil obtener la distribución *límite* de los términos de la serie variacional cuando el volumen de la muestra X crece ilimitadamente. Aquí sólo examinaremos un resultado que se deriva de los teoremas de continuidad.

Teorema 2. *Si* $a = \dfrac{k}{n+1} \to a_0 \in (0, 1)$ *cuando* $n \to \infty$, *entonces*

$$y_{(k)} = a + \frac{\sqrt{a_0(1-a_0)}}{\sqrt{n}}\,\xi_n,\ \xi_n \in \Phi_{0,1}.$$

Demostración. En virtud del teorema 1, $y_{(k)} \in \mathbf{B}_{k,\,n-k+1}$ y, por lo tanto, en virtud de las propiedades de la B-distribución, es válida la representación

$$y_{(k)} \underset{d}{=} \frac{\eta_1}{\eta_1 + \eta_2},\ \eta_j \in \mathbf{H}_{k_j},\ k_1 = 2k,\ k_2 = 2(n-k+1).$$

Pongamos, para comodidad, $a_1 = a$, $a_2 = 1 - a$, y supongamos que $a = a_0$ ha sido fijado. Entonces, evidentemente, $k_j/(n+1) = 2a_j$, $j = 1, 2$ y, en virtud de la propiedad de la distribución χ^2,

$$\eta_j = k_j + \sqrt{2k_j}\,.\ \xi_n^{(j)} \Rightarrow \xi^{(j)} \in \Phi_{0,1};$$

$$y_{(k)} = \frac{a_1 + \sqrt{\dfrac{a_1}{n+1}}\ \xi_n^{(1)}}{a_1 + a_2\sqrt{\dfrac{a_1}{n+1}}\ \xi_n^{(1)} + \sqrt{\dfrac{a_2}{n+1}}\ \xi_n^{(2)}}.$$

Nos queda utilizar el teorema de continuidad 1.5.3A para

$$H(t) = \frac{t_1}{t_1 + t_2},\ b_n = \frac{1}{\sqrt{n+1}},\ \eta_n^{(j)} = \sqrt{a_j}\,\xi_n^{(j)}.$$

Como η_j (y, por lo tanto, también $\xi_n^{(j)}$) son independientes y

$$\frac{\partial H}{\partial t_1} = \frac{t_2}{(t_1+t_2)^2},\quad \frac{\partial H}{\partial t_2} = -\frac{t_1}{(t_1+t_2)^2},$$

obtenemos

$$(y_{(k)} - a_1)\sqrt{n+1} \Rightarrow a_2\sqrt{a_1}\ \xi^{(1)} - a_1\sqrt{a_2}\ \xi^{(2)} \underset{d}{=} \sqrt{a_1 a_2}\ \xi,\ \xi \in \Phi_{0,1}.$$

Si a depende de n, entonces conviene utilizar la observación 1.5.1. ◁

Corolario 1. *Si* $a = k/(n+1) \to a_0 \in (0, 1)$ *y la función continua* F *es continuamente derivable en el punto* $\zeta_0 = F^{-1}(a_0)$ *(cuantila de orden* a_0*), entonces*

$$x_{(k)} = \zeta + \frac{\sqrt{a_0(1-a_0)}\ \xi_n}{f(\zeta_0)\sqrt{n}},\ \xi_n \in \Phi_{0,1}, \tag{11}$$

donde $\zeta = F^{-1}(a)$ *es una cuantila de orden a,* $f(x) = F'(x)$.

Esta afirmación se obtiene directamente del teorema de continuidad 1.5.3 (teniendo en cuenta la observación 1.5.1) si se utiliza la representación

$$x_{(k)} = F^{-1}(y_{(k)}) = F^{-1}\left(a + \sqrt{\frac{a_0(1-a_0)}{n}}\ \xi_n\right)$$

y el hecho de que $\dfrac{dF^{-1}(x)}{dx} = \dfrac{1}{f(F^{-1}(x))}$.

Observación 1. La afirmación (11) generaliza, de cierto modo, la afirmación del corolario 1.8.1. La misma también puede ser generalizada en otro sentido. Sea, para $x \to \zeta$,

$$|F(x) - F(\zeta)| \sim c\,|x - \zeta|^{\gamma},\ \gamma > 0.$$

Entonces es fácil ver que, cuando $y \to a$,

$$|F^{-1}(y) - F^{-1}(a)| \sim \left|\frac{y-a}{c}\right|^{1/\gamma}$$

y, por lo tanto,

$$(x_{(k)} - \zeta)n^{\frac{1}{2\gamma}} \Rightarrow (a_0(1-a_0))^{\frac{1}{2\gamma}}\,|\xi_c|^{\frac{1}{\gamma}}\,\operatorname{sign}\xi,\ \xi \in \Phi_{0,1} \tag{12}$$

Cuando $\gamma = 1$, $c = f(\zeta)$, de aquí se deduce (11).

10. Distribución de Cauchy $\mathbf{K}_{\alpha,\sigma}$ con parámetros (α, σ). Así se llama la distribución de densidad

$$k_{\alpha,\sigma}(x) = \frac{\sigma}{\pi\,|\sigma^2 + (x-\alpha)^2|} = \frac{1}{\pi\sigma} \cdot \frac{1}{1 + \left(\dfrac{x-\alpha}{\sigma}\right)^2}.$$

Al igual que en el caso de la ley normal, aquí los parámetros α y σ son, respectivamente, los parámetros de desplazamiento y de escala. La forma de la distribución $\mathbf{K}_{0,1}$ es muy semejante a la de $\Phi_{0,1}$, sin embargo, $k_{0,1}$, al igual que la densidad de la distribución de Student, tiene "colas mucho más gruesas" (o sea, un decrecimiento más lento cuando $|x| \to \infty$), así que la distribución $\mathbf{K}_{0,1}$ no tiene incluso una esperanza matemática finita. En [11] hemos señalado (véase el cap. 7) que las distribuciones $\mathbf{K}_{\alpha,\sigma}$, al igual que las distribuciones normales, poseen propiedad de estabilidad. La función característica $\chi_{0,1}(t)$ de la distribución $\mathbf{K}_{0,1}$ es igual a

$$\chi_{0,1}(t) = e^{-|t|},$$

por eso $\chi_{\alpha,\sigma}(t) = \exp\{i\alpha t - \sigma|t|\}$,

$$\chi_{\alpha_1,\ \sigma_1}(t)\chi_{\alpha_2,\ \sigma_2}(t) = \exp\{i(\alpha_1 + \alpha_2)t - (\sigma_1 + \sigma_2)\,|\,t\,|\,\},$$

así que la convolución de $\mathbf{K}_{\alpha_1,\ \sigma_1}$ y $\mathbf{K}_{\alpha_2,\ \sigma_2}$ es igual a $\mathbf{K}_{\alpha_1+\alpha_2,\ \sigma_1+\sigma_2}$. No es difícil ver que $\mathbf{K}_{0,1} = \mathbf{T}_1$.

En las aplicaciones se encuentran con frecuencia las funciones de diferente género de las variables aleatorias normalmente distribuidas. Una de ellas es la función exponencial con la cual está relacionada la llamada distribución lognormal.

11. Distribución lognormal $\mathbf{L}_{\alpha,\sigma^2}$. Diremos que $\eta \in \mathbf{L}_{\alpha,\sigma^2}$ si $\ln \eta \in \Phi_{\alpha,\sigma^2}$. En otros términos, $\eta = e^{\xi}$, donde $\xi \in \Phi_{\alpha,\sigma^2}$. De aquí se deduce que la distribución $\mathbf{L}_{\alpha,\sigma^2}$ está concentrada en el semieje positivo.

La densidad de $\eta \in \mathbf{L}_{\alpha,\sigma^2}$, en virtud de las fórmulas para la densidad de la función de la variable aleatoria (véase [11], p. 53), es igual a

$$\varphi_{\alpha,\sigma^2}(\ln x)x^{-1}.$$

Además, hallamos

$$\mathbf{M}\eta = \int e^{y}\,\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(y-\alpha)^2}{2\sigma^2}}\,dy = \exp\frac{(a+\sigma^2)^2 - \alpha}{2\sigma^2} \times$$

$$\times \int \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\alpha-\sigma^2)^2}{2\sigma^2}\right)dy = e^{\alpha+\sigma^2/2},$$

$$\mathbf{M}\eta^2 = \int e^2 y\,\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(y-\alpha)^2}{2\sigma^2}}\,dy = e^{2\alpha+2\sigma^2}.$$

12. Distribución degenerada. El símbolo $\mathbf{I}_a$ (ya hemos utilizado esta designación en el § 1.2) significará la distribución degenerada concentrada en el punto a.

En el caso general, cuando se examina una *familia arbitraria de distribuciones* que dependen del parámetro θ (escalar o vectorial), utilizaremos la designación $\mathbf{P}_\theta$. La propia familia se designará con el símbolo

$$\{\mathbf{P}_\theta\}_{\theta\in\Theta}$$

donde Θ es el conjunto de valores posibles del parámetro θ. Estas mismas designaciones se emplearán para las familias de distribuciones 1—12. Por ejemplo, $\{\Phi_{\alpha,1}\}_{\alpha\in R}$ significará la familia de todas las distribuciones normales con una varianza unitaria.

Las distribuciones 1—11 son absolutamente continuas con respecto a la medida de Lebesgue. Introduzcamos ahora las designaciones para tres distribuciones *discretas* bien conocidas (absolutamente continuas con respecto a la medida de cálculo $\mu(B)$:$\mu(B) = k$ si B contiene k puntos de valores enteros).

13. Distribución de Bernoulli $\mathbf{B}_p^n$. Según la definición, $\xi \in \mathbf{B}_p^n$ (n es un número entero, $p \in [0, 1]$) si

$$\mathbf{P}(\xi = k) = C_n^k p^k (1-p)^{n-k}, \quad 0 \leqslant k \leqslant n.$$

14. Distribución de Poisson Π_λ. Esta distribución se determina por medio de la igualdad

$$\Pi_\lambda(B) = \sum_{\substack{k \in B \\ k \geqslant 0}} \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0.$$

15. Distribución polinomial. Designaremos esta distribución por $\mathbf{B}_p^n$, donde $n > 0$ es un número entero, $p = (p_1, \ldots, p_r)$, $p_j \geqslant 0$, $\sum_{j=1}^{r} p_j = 1$. Para el vector aleatorio entero $\nu = \nu_1, \ldots, \nu_r)$ escribiremos $\nu \in \mathbf{B}_p^n$ si para $k = (k_1, \ldots, k_r)$, $k_j \geqslant 0$, $\sum_{j=1}^{r} k_j = n$ es válida la igualdad

$$\mathbf{P}(\nu = k) = \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r}.$$

La distribución $\mathbf{B}_p^n$ corresponde a la sucesión de n pruebas independientes, en cada una de las cuales se produce uno de r casos posibles incompatibles $A_1, \ldots, A_r$; entonces la probabilidad de que aparezca el caso A_j en una prueba es igual a p_j. Las coordenadas ν_j del vector ν significan las frecuencias de aparición de los sucesos A_j después de n pruebas (véase, por ejemplo, [11]). Es evidente que para cada $j = 1, \ldots, r$

$$\nu_j \in \mathbf{B}_{p_j}^n.$$

En el experimento ilustrado, el caso de la j-ésima prueba puede ser descrito por el vector de r-coordenada x_j, cuya $r-1$ coordenadas son iguales a cero, y una coordenada es igual a 1. El número de esta coordenada es el número del suceso que se produjo en la j-ésima prueba. Evidentemente que $\nu = \sum_{j=1}^{n} x_j$. Con respecto a la muestra X, formada por $x_1, \ldots, x_n$, nos será más cómodo escribir

$$X \in \mathbf{B}_p,$$

donde $\mathbf{B}_p = \mathbf{B}_p^1$. El espacio $\mathscr{X}$ para tal muestra es, por lo visto, finito y consta de r puntos. Si $p = (p_1, p_2)$, $p_1 + p_2 = 1$, obtendremos el esquema de Bernoulli, para el cual utilizaremos las mismas designaciones, identificando $\mathbf{B}(p_1, p_2)$ con $\mathbf{B}_{p_1} = \mathbf{B}_{p_1}^1$ (véase el subpárr.13). En el caso general, la distribución $\mathbf{B}_p$ depende, en realidad, solamente del parámetro de dimensión $r-1$ $(p_1, \ldots, p_{r-1})$, así que en vez del índice p se podría escribir $(p_1, \ldots, p_{r-1})$.

Muchas de las distribuciones examinadas más arriba, por ejemplo las

distribuciones $\Phi_{0,1}$, $\mathbf{H}_k$, $\mathbf{F}_{k_1,\ k_2}$, $\mathbf{T}_k$, Π_λ, están tabuladas en los manuales de estadística matemática y se ofrecen en tablas especiales (véase, por ejemplo, [8]).

§ 3. Estimación puntual. Método principal de obtención de estimaciones. Conciliabilidad. Normalidad asintótica

1. Método de sustitución. Conciliabilidad. En el § 1 hemos introducido el concepto de estimación. Formalmente, estimación es lo mismo que estadística, o sea, toda función medible θ^* de una muestra. No formalmente, el sentido que se le da a este término consiste en que llamamos estimaciones θ^* sólo a las estadísticas que deben utilizarse en vez del parámetro desconocido θ. Con otras palabras, θ^* es cierta aproximación para θ, basada en la muestra. La magnitud θ^* también se denomina estimación *puntual* para θ, a distinción de las estimaciones *por intervalo* que serán examinadas más adelante.

La representación de una estimación presupone, de ordinario, la representación de funciones (de la muestra X_n) definidas para todos los valores posibles de n. Por eso, en lo sucesivo el término "estimación" significará la familia de estadísticas $\theta^* = \theta_n^*(X_n)$ definidas para todas los $n = 1, 2, \ldots$, donde θ^* es la función sobre $\mathscr{X}^n$, o bien, que es lo mismo, una función $\theta^* = \theta^*(n, X_\infty)$ definida en el producto del conjunto de números enteros y $\mathscr{X}^\infty$.

De acuerdo con el § 1, consideraremos que en el planteamiento del problema de estimación está definido el conjunto Θ de los posibles valores del parámetro θ y la familia $\mathscr{P}$ de las posibles distribuciones $\mathbf{P}$ de la muestra X (que pueden ser, digamos, sólo las distribuciones normales $\Phi_{\alpha,\ 1}$ o las distribuciones de Poisson Π_λ para las cuales es preciso estimar los parámetros desconocidos α, λ). Si faltan cualesquiera limitaciones para θ (o para $\mathbf{P}$), entonces podemos considerar que $\theta\mathscr{P}$ coincide con el espacio euclidiano de dimensión correspondiente (con el conjunto de todas las distribuciones).

Si para designar el parámetro, en vez de θ se utiliza otra letra cualquiera, por ejemplo λ, las estimaciones de este parámetro se designarán del mismo modo: añadiendo a λ el índice superior en forma de asterisco. Por ejemplo, para el parámetro α de la ley normal es natural examinar la estimación

$$\alpha^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Los momentos muestrales que se utilizan para la estimación

$$\mathbf{M}x_1 = \int x\mathbf{P}(dx) \text{ y } \mathbf{D}x_1 = \int (x - \mathbf{M}x_1)^2\mathbf{P}(dx)$$

tienen sus designaciones especiales tradicionales

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

Ya hemos señalado que para el parámetro dado se pueden indicar varias estimaciones, tantas como se quiera, y antes de examinar de qué modo en cada situación concreta conviene comparar sus cualidades, fijaremos la atención en ciertos métodos "regulares" generales de su construcción.

Estos métodos agrupan en sí los enfoques más racionales del problema de estimación y posteriormente nos permitirán obtener las mejores estimaciones en uno u otro sentido.

Casi todos los procedimientos de estimación se basan en el siguiente método principal, que podría llamarse *método de sustitución de la distribución empírica* (o simplemente *método de sustitución*).

Sea $X_n \in \mathbf{P}$ y representemos el parámetro desconocido θ en forma de cierta funcional G de la distribución $\mathbf{P}$:

$$\theta = G(\mathbf{P}) \cdot$$

Supongamos, luego, que $\mathbf{P}_n^*$ significa, como antes, la distribución empírica. Entonces, el método de sustitución prescribe que en calidad de estimación θ^* se tome la función

$$\theta^* = G(\mathbf{P}_n^*).$$

Tales estimaciones serán llamadas *estimaciones por el método de sustitución* o simplemente *estimaciones de sustitución.*

La funcional G se da, a veces, en forma implícita como solución de cierta ecuación $H(\theta, \mathbf{P}) = 0$, resoluble con respecto a θ. En este caso, en consonancia con la definición principal, llamaremos estimaciones de sustitución a toda solución de la ecuación $H(\theta, \mathbf{P}_n^*) = 0$.

Si se sabe que el conjunto de los posibles valores del parámetro $\theta \in \in R^k$ está limitado por el dominio Θ de R^k, el cual no coincide con R^k, esta información se puede tener en cuenta al construir las estimaciones de sustitución. Admitamos que el domino Θ está cerrado y sea $\mathscr{P}$ el conjunto de las posibles distribuciones de la muestra X, $\Theta = \{G(\mathbf{P})\}_{\mathbf{P}\in\mathscr{P}}$ Definamos la funcional $G_1(\mathbf{P})$ para $\mathbf{P}$ arbitraria, como el valor de $t \in \Theta$ para el que se alcanza

$$\min_{t\in\Theta} |t - G(\mathbf{P})| = |G_1(\mathbf{P}) - G(\mathbf{P})|, \tag{1}$$

así que $G_1(\mathbf{P})$ es el punto de Θ más próximo a $G(\mathbf{P})$. Como $G_1(\mathbf{P}) = G(\mathbf{P}) = \theta$, si $\mathbf{P} \in \mathscr{P}$, entonces la estimación

$$\theta^* = G_1(\mathbf{P}_n^*), \tag{2}$$

junto con $G(\mathbf{P}_n^*)$, será la estimación de sustitución, con la particularidad de que el conjunto de los posibles valores de θ^* pertenecerá a Θ.

En cuanto a las estimaciones (1) y (2) diremos que se han obtenido debido a la *contracción del método de sustitución.*

Supongamos, por ejemplo, que se estima el parámetro α de la distribución normal $\Phi_{\alpha,1}$ y que sabemos de antemano que $\alpha \in [0, 1]$. Entonces puede resultar que la estimación $\alpha^* = \overline{x} \notin [0, 1]$ (evidentemente que $\overline{x} = \int t dF_n^*(t)$ es la estimación de sustitución). La contracción del método de sustitución recomienda en calidad de estimación tomar el punto [0, 1] más próximo a $\overline{x}$.

Señalemos ahora, que en la forma enunciada, el método de sustitución no siempre tiene sentido. El hecho consiste en que la funcional inicial G puede resultar no definida sobre el conjunto de distribuciones empíricas. Supongamos, por ejemplo, que es sabido de antemano que la distribución $\mathbf{P}$ pertenece a la clase $\mathscr{P}$ de distribuciones absolutamente continuas con respecto a la medida de Lebesgue, así que cada $\mathbf{P} \in \mathscr{P}$ tiene una densidad igual a f.

Pero a nosotros nos interesa el valor de

$$\theta = G(\mathbf{P}) = \int f^2(x)dx = \int \left(\frac{d\mathbf{P}}{dx}\right)^2 dx.$$

Está claro que en este caso $G(\mathbf{P}_n^*)$ no tiene sentido, ya que $\mathbf{P}_n^*$ es una distribución discreta. En tales casos el método de sustitución siempre puede ser modificado naturalmente de manera que conserve su esencia. En el ejemplo citado, donde $G(\mathbf{P})$ es la funcional de la densidad f, conviene, en calidad de θ^*, examinar, de acuerdo con el método de sustitución, el valor de $G(\mathbf{P}_n^{**})$, donde $\mathbf{P}_n^{**}$ es la distribución empírica suavizada (véase el § 1.10) que asegura la convergencia de la densidad empírica hacia $f(x)$.

También puede resultar que en algunos casos $G(\mathbf{P}_n^*)$ tenga sentido no para todas las X_n, sino sólo para $X_n \in A_n$, donde $\mathbf{P}(X_n \in A_n) \to 1$ cuando $n \to \infty$. Esta circunstancia no tendrá ninguna importancia en cuanto a la esencia de la exposición ulterior del material, y para precisar podemos poner $G(\mathbf{P}_n^*) = 0$ para $X_n \notin A_n$. En este párrafo, para simplificar, estimaremos que $G(\mathbf{P}_n^*)$ tiene sentido para todas $X_n \in \mathscr{X}^n$, y que θ^* es una variable aleatoria, o sea, que la función $G(\mathbf{P}_n^*)$ realiza la aplicación medible de $\mathscr{X}^n$ en R^k, donde k es la dimensión de θ.

El principio de sustitución es un enfoque muy natural del problema, puesto que, como ya sabemos, la distribución $\mathbf{P}_n^*$ se aproxima ilimitadamente a $\mathbf{P}$ a medida que crece n.

Sea $X_n = |X_\infty|_n$.

Definición 1. La estimación $\theta^* = \theta_n^*(X_n)$ (o la sucesión $\theta_n^*(X)$) se llama

conciliable si

$$\theta^* \xrightarrow[P]{} \theta$$

cuando $n \to \infty$.

La estimación θ^* se denomina *fuertemente conciliable* si, para $n \to \infty$,

$$\theta^* \xrightarrow[\text{c.s}]{} \theta$$

Sea F, como siempre, la función de distribución correspondiente a $\mathbf{P}$.

Teorema 1. *Supongamos que* $\theta = G(\mathbf{P})$ *y que la funcional* G *pertenece a una de las dos clases, o que es representable en la forma*

$$G(\mathbf{P}) = h\left(\int g(x)dF(x)\right), \tag{I}$$

donde h *es una función continua en el punto* $a = \int g(x)dF_0(x)$ *(funcional de tipo I), o representable en la forma*

$$G(\mathbf{P}) = G_1(F), \tag{II}$$

donde la funcional G_1 *es continua en el punto* F_0, *en la métrica uniforme (funcional de tipo* II*). Entonces, si* $X \in F_0$, $\theta^* = G(\mathbf{P}_n^*)$ *es una estimación fuertemente conciliable:*

$$\theta^* \xrightarrow[\text{c.s.}]{} \theta.$$

La afirmación de este teorema se deduce directamente del teorema 1.4.1.

2. Normalidad asintótica. Caso unidimensional.

Definición 2. La estimación θ^* del parámetro θ se llama *asintóticamente normal (a.n.) con coeficiente* $\sigma^2 \geqslant 0$, si $(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2}$.

La última relación también puede leerse del modo siguiente: la estimación θ^* a.n. con los parámetros $(\theta, \sigma^2/n)$.

Supongamos que θ^* es la estimación de sustitución del parámetro $\theta = G(\mathbf{P})$ y que se cumple (I), o sea, que

$$\theta^* = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right) \tag{3}$$

es una estadística de tipo 1. Entonces, de los resultados del § 1.7 se deduce la afirmación siguiente. Supongamos que θ es un parámetro escalar, y g, una función escalar.

Teorema 2. *Sea* $X \in F_0$, h *derivable en el punto* $a = \int g(x)dF_0(x)$, $0 < |h'(a)| < \infty$, $\int g^2(x)dF_0(x) < \infty$. *Entonces* θ^* *es la estimación a.n. con coeficiente*

$$\sigma^2 = [h'(a)]^2 \int (g(x) - a)^2 dF_0(x).$$

Los ejemplos examinados en el § 1.7 también pueden utilizarse como ilustraciones de este teorema, ya que las estadísticas examinadas en ellos se utilizan en calidad de estimaciones.

Análogamente podríamos, utilizando los resultados del § 1.8, obtener las condiciones de normalidad asintótica de las estimaciones que son estadísticas de tipo II. El lector puede obtener las afirmaciones necesarias, utilizando el teorema 1.8.1 sin cualesquiera modificaciones, pero exigiendo, no obstante, que en su enunciación se cumpla $k = 1$, y que la derivada g sea tal que $g(F_0, w^\circ) \in \Phi_{0,\sigma^2}$.

3. Normalidad asintótica. Caso de parámetro multidimensional.

Definición 2A. La estimación $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ se denomina *estimación a.n* $\theta = (\theta_1, \ldots, \theta_k)$ *con matriz* σ^2, *si*

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2}, \tag{4}$$

donde Φ_{a,σ^2} es la distribución normal k-dimensional con vector nulo de las esperanzas matemáticas y con matriz de segundos momentos $\sigma^2 = \|\sigma_{ij}\|$. La densidad de esta distribución es igual (véase el § 2) a

$$\Phi_{0,\sigma^2}(x) = \frac{\sqrt{|A|}}{(2\pi)^{k/2}} e^{-\frac{1}{2} xAx^T},$$

donde A es una matriz inversa a σ^2, $x = (x_1, \ldots, x_k)$.

Si θ^* es la estimación de la sustitución y la misma es una estadística de tipo I (o sea, representable en forma de (3), donde g, hablando en general, junto con θ^* y h, es una función vectorial), entonces, para determinar las condiciones de normalidad asintótica se puede utilizar el teorema 1.7.1A y la observación a él. En este caso obtenemos la afirmación siguiente.

Teorema 2A. *Supongamos que* $\theta^* \in R^k$ *se define por la igualdad* (1), *donde* $g = (g_1, \ldots, g_s) \in R^s$, *y la función vectorial* $h(t) = (h_1(t), \ldots, h_k(t))$, $t = (t_1, \ldots, t_s)$ *tiene en el punto* $a = (a_1, \ldots, a_s)$, $a_j = \int g_j(x) dF_0(x)$ *las derivadas parciales* $\frac{\partial h_l}{\partial t_j}(a)$, $l = 1, \ldots, k$, $j = 1, \ldots, s$. *Entonces, si* $X \in F_0$,

$$(\theta^* - \theta)\sqrt{n} \Rightarrow \xi H^T,$$

donde $\xi = (\xi_1, \ldots, \xi_s) \in \Phi_{0,d^2}$ *es el vector normalmente distribuido, con la media nula y la matriz de segundos momentos* $d^2 = \|d_{ij}\|$, $d_{ij} = \mathbf{M}(g_i(x_1) - a_i)(g_j(x_1) - a_j)$, $i, j = 1, \ldots, s$; $H = \|h_{ij}\|$ *es una matriz de dimensión* $k \times s$, *con los elementos* $h_{ij} = \frac{\partial h_i}{\partial t_j}(a)$, $i = 1, \ldots, k$; $j = 1, \ldots, s$.

Esto significa, a su vez, que al cumplirse las condiciones del teorema 2A, θ^* es una estimación a.n. con matriz $\sigma^2 = Hd^2H^T = \mathbf{M}H\xi^T\xi H^T$. Cabe

señalar que las matrices σ^2 y d^2 aquí tienen, hablando en general, dimensiones diferentes (k y s).

§ 4. Realización del método de sustitución en el caso paramétrico. Método de momentos

Sea $X \in \mathbf{P}_\theta$, donde $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ es la familia de distribuciones $\mathbf{P}_\theta$ que ya conocemos y que dependen del parámetro θ. En nuestras investigaciones, el parámetro desconocido θ del conjunto Θ puede ser tanto escalar como vectorial. Por ejemplo, si $X \in \Phi_{\alpha,\sigma^2}$, entonces $\theta = (\alpha, \sigma^2)$ es bidimensional, y el conjunto Θ puede ser tanto un semiplano $\{-\infty < \alpha < \infty, \sigma \geqslant 0\}$ como cualquier parte de éste.

La esperanza matemática y la varianza de la estadística $S = S(X)$ en función de la distribución $\mathbf{P}_\theta$ serán designadas por $\mathbf{M}_\theta S$ y $\mathbf{D}_\theta S$, respectivamente.

Más adelante examinaremos algunos métodos de estimación, cada uno de los cuales puede interpretarse como la realización del principio de sustitución de una distribución empírica.

1. Método de momentos. Caso unidimensional. Escojamos $g(x)$ de tal modo que la función

$$m(\theta) = \mathbf{M}_\theta g(\mathrm{x}_1) = \int g(x)\mathbf{P}_\theta(dx) \tag{1}$$

sea monótona y continua. El campo $m(\Theta)$ de valores $m(\theta)$, $\theta \in \Theta$ tiene la misma "naturaleza" que Θ. Si, por ejemplo, Θ es un segmento del eje real, $m(\Theta)$ también será un segmento.

Es evidente que la ecuación $m(\theta) = t$ es unívoca y continuamente resoluble en el campo $m(\Theta)$ respecto a θ: $\theta = m^{-1}(t)$, y que (1) se puede escribir del modo equivalente en la forma

$$\theta = m^{-1}\Big(\int g(x)\mathbf{P}_\theta(dx)\Big). \tag{2}$$

Supongamos simplemente, que

$$\overline{g} \equiv \int g(x)d\mathbf{P}_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} g(\mathrm{x}_i) \in m(\Theta)$$

para todas $X \in \mathscr{X}^n$.

Definición 1. Se llama estimación por el *método de momentos* la estimación

$$\theta^* = m^{-1}(\overline{g}).$$

Si $\overline{g} \notin m(\Theta)$, se puede poner, conforme a (3.1) y (3.2),

$$\theta^* = m^{-1}(\overline{g}_0),$$

donde $\bar{g}_0 \in m(\Theta)$ es el punto de $m(\Theta)$ más próximo a $\bar{g}$.

No es difícil darse cuenta que esto constituye la estimación con arreglo al principio de sustitución. La elección de la función $m(\theta)$ nos ha permitido expresar θ en forma de la funcional (2). También está claro que la estimación (3) es una estadística de tipo I, así que, en virtud del teorema 3.1, las estimaciones conforme al método de momentos serán fuertemente conciliables. Si además, la función m es derivable en el punto θ, $\int g^2(x)\mathbf{P}_\theta(dx) < \infty$, entonces, según el teorema 3.2, la estimación con arreglo al método de momentos será a.n. con coeficiente $(m'(\theta))^{-2}\mathbf{D}_\theta g(\mathrm{x}_1)$.

El método de momentos fue propuesto por C. Pearson (en forma algo más particular) e históricamente es el primer método regular para construir estimaciones.

La propia denominación de "método de momentos" se debe al hecho de que su esencia consiste en igualar entre sí los momentos "teóricos" y empíricos (esperanzas matemáticas) de la magnitud $g(\mathrm{x}_1)$: pues la estimación (3) no es otra cosa sino la solución de la ecuación

$$m(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(\mathrm{x}_i). \tag{4}$$

También se puede añadir que en calidad de $g(x)$ se elige con frecuencia la función $g(x) = x$ o bien $g(x) = x^k$, $k > 1$, así que nuestra ecuación se convierte en ecuación para momentos ordinarios.

La igualdad (4) también puede considerarse como el resultado de la igualación del valor medio de la magnitud $g(\mathrm{x}_1)$ "en el espacio", a su valor medio "en el tiempo".

El carácter no únivoco del método de momentos, así como de todo el principio de sustitución, aquí se manifiesta sobre todo bien: pues casi nada limita la elección de la función $g(x)$.

Ejemplo 1. Supongamos que $X \in \mathbf{\Gamma}_{\alpha,1}$ y que α se desconoce. Construyamos las estimaciones conforme al método de momentos con dos funciones elementales $g(x)$:$g_1(x) = x$ y $g_2(x) = x^2$. Son válidas las igualdades siguientes (véase el punto 5 del § 2)):

$$m_1(\alpha) = \mathbf{M}_\alpha g_1(\mathrm{x}_1) = \int_0^\infty x\Gamma_{\alpha,1}(dx) = 1/\alpha,$$

$$m_2(\alpha) = \mathbf{M}_\alpha g_2(\mathrm{x}_1) = \int_0^\infty x^2\Gamma_{\alpha,1}(dx) = 2/\alpha^2.$$

Resolviendo las ecuaciones $m_1(\alpha) = \bar{\mathrm{x}}$, $m_2(\alpha) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, obtenemos

las estimaciones según el método de momentos

$$\alpha^* = (\bar{x})^{-1} \text{ y } \alpha^{**} = \left(\frac{1}{2n}\sum_{i=1}^{n} x_i^2\right)^{-1/2}, \tag{5}$$

Estas dos estimaciones son estadísticas de tipo I y podemos describir sus propiedades asintóticas. A base de las igualdades (2.4) obtenemos

$$D_\alpha g_1(x_1) = \mathbf{D}_\alpha x_1 = 1/\alpha^2, \ \mathbf{D}_\alpha g_2(x_1) = \mathbf{D}_\alpha x_1^2 = 20/\alpha^4.$$

En vista de que para la primera estimación, $m_1'(\alpha) = -1/\alpha^2$, y para la segunda, $m_2'(\alpha) = -4/\alpha^3$, a base de los teoremas 3.1, 3.2 obtenemos que ambas estimaciones α^* y α^{**} son fuertemente conciliables y a.n. con coeficientes, respectivamente,

$$\frac{1}{\alpha^2} \cdot \alpha^4 = \alpha^2, \ \frac{20}{\alpha^4} \cdot \frac{\alpha^2}{16} = \frac{5}{4}\alpha^2.$$

Evidentemente, conviene dar preferencia a α^*, ya que su "dispersión", en caso de grandes valores de n alrededor del valor verdadero de α, que se mide con arreglo a la varianza de la distribución límite, es menor que la "dispersión" para α^{**}.

2. Método de momentos. Caso multidimensional. De un modo completamente análogo se examina el caso cuando θ es un parámetro multidimensional.

Supongamos, como antes, que k es la dimensión de θ. Elijamos la función vectorial $g(x) = (g_1(x), \ldots, g_k(x))$ de modo que la ecuación

$$m(\theta) = t,$$

donde $t = (t_1, \ldots, t_k)$, $m(\theta) = (m_1(\theta), \ldots, m_k(\theta))$,

$$m_j(\theta) = \mathbf{M}_\theta g_j(x_1) = \int g_j(x)\mathbf{P}_\theta(dx),$$

sea unívoca y continuamente resoluble con respecto a $\theta = m^{-1}(t)$ en el campo $m(\Theta)$ de valores $m(\theta)$, $\theta \in \Theta$. Admitamos simplemente, que el vector

$$\bar{g} = \left(\frac{1}{n}\sum_{i=1}^{n} g_1(x_i), \ \ldots, \ \frac{1}{n}\sum_{i=1}^{n} g_k(x_i)\right)$$

pertenece al campo $m(\Theta)$ de todas $X \in \mathscr{X}^n$.

Definición 1A. La estimación $\theta^* = m^{-1}(\bar{g})$ se llama *estimación por el método de momentos.*

Como antes, del teorema 3.1 se deduce que tales estimaciones θ^* serán fuertemente conciliables.

Para que tenga lugar θ^* a.n. es necesario exigir adicionalmente que la función m sea derivable, $\int g_i^2(x)\mathbf{P}_\theta(dx) < \infty$. La afirmación acerca de la distribución límite de θ^* se obtiene fácilmente con ayuda del teorema 3.2A.

Ejemplo 2. Examinemos en calidad de $\{\mathbf{P}_\theta\}$ la familia de distribuciones normales Φ_{α,σ^2}. Suponiendo $g_1(x) = x$, $g_2(x) = x^2$, obtenemos las ecuaciones siguientes para el método de momentos:

$$\alpha = \bar{x},\ \sigma^2 + \alpha^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

cuya solución es

$$\alpha^* = \bar{x},\ (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = S^2.$$

Proponemos al lector, en calidad de ejercicio, hallar, basándose en el método de momentos, las estimaciones para todas las familias paramétricas expuestas en el § 2.

3. Método generalizado de momentos. Es posible la siguiente generalización del método de momentos, la cual amplía considerablemente la clase de estimaciones antes examinada. Limitémonos simplemente al caso del parámetro unidimensional θ. Examinemos la función de dos variables $g(x, \theta)$ y supongamos que para toda distribución $\mathbf{P}$ la ecuación

$$\int g(x, \theta)\mathbf{P}(dx) = \int g(x, \theta)\mathbf{P}_\theta(dx) \tag{6}$$

es resoluble con respecto a $\theta = G(\mathbf{P})$, de modo que la última igualdad, junto con (6), se convierta en la identidad $\theta = G(\mathbf{P}_\theta)$ cuando $\mathbf{P} = \mathbf{P}_\theta$.

Llamaremos *estimación por el método generalizado de momentos*, la estimación

$$\theta^* = G(\mathbf{P}_n^*).$$

Es evidente que, al igual que las estimaciones por el método de momentos, éstas son estimaciones de sustitución. La investigación de las propiedades de tales estimaciones es más difícil. De esto nos convenceremos en los párrafos sucesivos, puesto que resultará que una de las estimaciones de sustitución que estudiaremos detalladamente será la estimación por el método generalizado de momentos.

§ 5*. Método de distancia mínima

El método indicado en el título, al igual que el de momentos, es la realización del principio de sustitución y consiste en lo siguiente. Examinemos cualquier funcional de dos distribuciones $d(\mathbf{P}, \mathbf{Q})$, la cual posee la propie-

dad consistente en que como función de $\mathbf{Q}$ dicha funcional alcanza su valor mínimo cuando $\mathbf{Q} = \mathbf{P}$ y $d(\mathbf{P}, \mathbf{Q}) > d(\mathbf{P}, \mathbf{P})$ cuando $\mathbf{Q} \neq \mathbf{P}$. Vamos a considerar la magnitud $d(\mathbf{P}, \mathbf{Q})$ (o bien $d(\mathbf{P}, \mathbf{Q}) - d(\mathbf{P}, \mathbf{P})$) como la "distancia" entre $\mathbf{Q}$ y $\mathbf{P}$, de modo que $\mathbf{P}$ se pueda determinar como el valor de $\mathbf{Q}$ con el que $d(\mathbf{P}, \mathbf{Q})$ alcanza su valor mínimo.

Supongamos ahora que $X \in \mathbf{P}$, $\mathbf{P}$ se desconoce y pertenece a la familia $\mathscr{P}$. Designemos por $(\mathbf{Q})_{\mathscr{P}}$ la distribución de $\mathscr{P}$ inmediata a la distribución $\mathbf{Q}$ en sentido de la distancia d, y supongamos que ella existe:

$$d((\mathbf{Q})_{\mathscr{P}}, \mathbf{Q}) = \min_{\Pi \in \mathscr{P}} d(\Pi, \mathbf{Q}),$$

así que $(\mathbf{Q})_{\mathscr{P}} = \mathbf{Q}$ si $\mathbf{Q} \in \mathscr{P}$.

Definición 1. Se llama *estimación de la distribución* $\mathbf{P}$ *conforme al valor mínimo de la distancia* d, la distribución $\mathbf{P}^* = (\mathbf{P}_n^*)_{\mathscr{P}} \in \mathscr{P}$, donde $\mathbf{P}_n^*$ es, como antes, la distribución empírica.

Ahora bien, cuando $\Pi = \mathbf{P}^* = (\mathbf{P}_n^*)_{\mathscr{P}}$ se minimiza $d(\Pi, \mathbf{P}_n^*)$. Si $\mathscr{P}$ coincide con el conjunto de todas las distribuciones, es evidente que $\mathbf{P}^* = \mathbf{P}_n^*$.

Supongamos ahora que $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta \in \Theta}$ es una familia paramétrica que satisface la condición siguiente:

A_0 $\qquad \mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2}$ cuando $\theta_1 \neq \theta_2$.

En este caso la aplicación de $\theta \to \mathbf{P}_\theta$ es biunívoca, por eso la distribución $\mathbf{P} \in \mathscr{P}$ permite restablecer únicamente el parámetro θ con el que $\mathbf{P} = \mathbf{P}_\theta$. Este hecho también puede expresarse de otra manera: existe la funcional G definida sobre $\mathscr{P}$ de tal modo que $\theta = G(\mathbf{P}_\theta)$.

Introduzcamos en este planteamiento la funcional $G_1(\mathbf{Q}) = G((\mathbf{Q})_{\mathscr{P}})$ que es, evidentemente, el valor de $\theta \in \Theta$ con el que $\mathbf{P}_\theta$ será la distribución inmediata a $\mathbf{Q}$ en sentido de la distancia d, así que

$$G_1(\mathbf{P}_\theta) = G(\mathbf{P}_\theta) = \theta. \tag{1}$$

Definición 2. La estimación $\theta^* = G_1(\mathbf{P}_n^*)$ se denomina *estimación del parámetro* θ *por el valor mínimo de la distancia* d.

En otros términos, θ^* es el valor de Θ con el que

$$d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*) = \inf_{\theta \in \Theta} d(\mathbf{P}_\theta, \mathbf{P}_n^*).$$

Es evidente que aquí otra vez tropezamos con el principio de sustitución. Esto se deduce de las definiciones y de (1). Claro está que la distancia d y la familia $\mathscr{P} = \{\mathbf{P}_\theta\}$ deben poseer propiedades capaces de asegurar la mensurabilidad de la aplicación de $\mathscr{X}^n$ en R^k, que se realiza mediante la funcional $G_1(\mathbf{P}_n^*)$, de modo que θ^* sea una variable aleatoria.

Ahora señalemos que *en el caso paramétrico, al cumplirse la condición*

(A_0), *la contracción del método de sustitución* (véanse (3.1) y (3.2)) *y el método de distancia mínima proporcionan la misma clase de estimaciones.*

En efecto, ya sabemos que las estimaciones de distancia mínimas θ^* son las estimaciones por el método de sustitución, en este caso $\theta^* \in \Theta$. Supongamos ahora que θ^* es la estimación por el método de sustitución $\theta^* = G(\mathbf{P}_n^*)$, donde $G(\mathbf{P}_\theta) = \theta$, $\theta^* \in \Theta$. Determinemos la distancia $d(\mathbf{P}, \mathbf{Q}) = |G(\mathbf{P}) - G(\mathbf{Q})|$. Entonces, evidentemente, para $\theta = \theta^*$ se alcanza

$$\inf_{\theta \in \Theta} d(\mathbf{P}_\theta, \mathbf{P}_n^*) = \inf_{\theta \in \Theta} |G(\mathbf{P}_\theta) - G(\mathbf{P}_n^*)| = \inf_{\theta \in \Theta} |\theta - G(\mathbf{P}_n^*)| = 0.$$

También se puede notar que el método de momentos es mucho más estrecho que el de sustitución, puesto que es evidente que no cada funcional G tal que $G(\mathbf{P}_\theta) = \theta$, admite la representación de la forma

$$G(\mathbf{P}_\theta) = m^{-1}\left(\int g(x)\mathbf{P}_\theta(dx)\right).$$

Volvamos a las estimaciones de distancia mínima. Está claro que se pueden señalar muchas distancias "racionales" d que pueden utilizarse para construir las estimaciones. Podríamos, en calidad de d, tomar la distancia

$$d(\mathbf{P}, \mathbf{Q}) = \sup_x |F_P(x) - F_Q(x)|$$

o bien

$$d(\mathbf{P}, \mathbf{Q}) = \int (F_P(x) - F_Q(x))^2 dF_Q(x),$$

donde $F_P(x)$ es la función de distribución que corresponde a la distribución $\mathbf{P}$. Aquí serán estimaciones θ^* por la distancia mínima los valores de θ con los que se alcanza, respectivamente,

$$\inf_\theta \sup_x |F_{P_\theta}(x) - F_n^*(x)|, \tag{2}$$

$$\inf_\theta \int (F_{P_\theta}(x) - F_n^*(x))^2 dF_n^*(x) = \inf_\theta \frac{1}{n} \sum_{k=1}^{n} \left(F_{P_\theta}(x_{(k)}) - \frac{k-1}{n}\right)^2.$$

En algunos problemas (compárese esto con [48]) se utilizan las llamadas estimaciones conforme al valor mínimo de χ^2 (ji-cuadrado). Son las estimaciones con arreglo al valor mínimo de la distancia

$$d(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{r} \frac{(\mathbf{P}(\Delta_i) - \mathbf{Q}(\Delta_i))^2}{\mathbf{P}(\Delta_i)},$$

donde $\Delta_1, \ldots, \Delta_r$ es la partición de R (o bien de R^m si x_j son m-dimensionales) en $r < \infty$ intervalos, así que $\bigcup_{i=1}^{r} \Delta_i = R$. Ahora bien, la

estimación θ^* conforme al valor mínimo de χ^2 es el valor de θ con el que se minimiza

$$n\sum_{i=1}^{r}\frac{(\mathbf{P}_\theta(\Delta_i)-\nu_i/n)^2}{\mathbf{P}_\theta(\Delta_i)}=\sum_{i=1}^{r}\frac{(n\mathbf{P}_\theta(\Delta_i)-\nu_i)^2}{n\mathbf{P}_\theta(\Delta_i)}. \tag{3}$$

Aquí $\nu_i = n\mathbf{P}_n^*(\Delta_i)$ es el número de observaciones x_j que adquirieron los valores del intervalo Δ_i. La estadística en el segundo miembro (3) es la estadística χ^2 que ya conocemos, de aquí precisamente procede la denominación de dicha estimación.

Más adelante veremos que existe tal funcional G, $\theta = G(\mathbf{P}_\theta)$ con la que las estimaciones según el principio de sustitución, llamadas estimaciones de verosimilitud máxima, serán las mejores en cierto sentido. En virtud de esta circunstancia, las estimaciones examinadas en este párrafo no tienen, hablando en general, mucha aplicación y por eso no merece la pena detenerse más en ellas.

§ 6. Método de verosimilitud máxima

Otra vez supongamos que $\mathscr{P}$ es una familia paramétrica $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$. En lo sucesivo, con arreglo a esta familia admitiremos, por doquier donde sea necesario, que está cumplida la condición

$$(A_0) \qquad \mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2} \quad \text{cuando} \quad \theta_1 \neq \theta_2,$$

así como la condición siguiente, que llamaremos condición (A_μ).

(A_μ): *en el espacio de fase* $(\mathscr{X}\mathfrak{B}_{\mathscr{X}})$ *existe una medida* σ*-finita* μ *tal que todas las distribuciones* $\mathbf{P}_\theta \in \mathscr{P}$ *tienen, respecto a esta medida, la densidad* $f_\theta(x) = \dfrac{d\mathbf{P}_\theta}{d\mu}(x)$, *así que*

$$\mathbf{P}_\theta(B) = \int_B f_\theta(x)\mu(dx).$$

En este caso se dice que la medida μ *domina* las distribuciones $\mathbf{P}_\theta$.

Todas las familias de distribuciones examinadas en el § 2 satisfacen, evidentemente, las condiciones (A_0) y (A_μ). Para ciertas distribuciones, en calidad de μ es necesario adoptar la medida de Lebesgue (distribuciones absolutamente continuas), y para otras, la medida de cálculo (distribuciones discretas). La medida de cálculo μ se define así: $\mu(B) = k$, donde k es el número de puntos con coordenadas de valores enteros pertenecientes a B.

A las primeras pertenecen las distribuciones normal Φ_{α,σ^2}, lognormal $\mathbf{L}_{\alpha,\sigma^2}$, las distribuciones Γ y B, la distribución uniforme, la distribución

de Cauchy y las distribuciones de Student y de Fisher, y a las segundas, las distribuciones de Bernoulli y Poisson, así como las distribuciones degeneradas en cero y polinomiales. La forma de densidades $f_\theta(x)$ de estas distribuciones se da en el § 2. En el caso discreto (cuando μ es la medida de cálculo), la densidad $f_\theta(x)$ coincide con la probabilidad $\mathbf{P}_\theta(\{x\})$ del suceso $\{x_1 = x\}$; aquí $\{x\}$ significa un conjunto compuesto por un solo punto x. También cabe señalar que, por ejemplo, la distribución normal Φ_{α,σ^2} y la distribución de Poisson son recíprocamente singulares. En vez de la medida de Lebesgue y la medida de cálculo también podríamos tomar otras medidas, por ejemplo, la distribución normal $\Phi_{0,1}$ y la distribución de Poisson Π_1, respectivamente. Sin embargo, en este caso las densidades $f_\theta(x)$ serán, evidentemente, otras. Proponemos que las halle el propio lector. Los ejemplos citados más arriba se referían al caso $\mathscr{X} = R$ o $\mathscr{X} = R^m$, $m > 1$. En un espacio de fase arbitrario $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$, la naturaleza de la medida μ puede ser más compleja.

La introducción de la condición $(A_\mu$ es cómoda, ante todo, por el hecho de que posteriormente nos permitirá examinar, desde un punto de vista único, dos tipos de distribuciones que son las más importantes en las aplicaciones: absolutamente continuas y discretas. Desde el punto de vista de la condición (A_μ), entre dichas distribuciones no hay ninguna diferencia cualitativa. Además, deja de ser importante la dimensión del espacio de fase $\mathscr{X}$.

Convengamos en escribir

$$f(x) = g(x) \quad \text{c.d. } [\mu]$$

si existe un conjunto A, $\mu(A) = 0$ tal que $f(x) = g(x)$ para todos $x \notin A$. Es evidente que $f(x) = g(x)$ c.s. $[\mu]$ si y sólo si

$$\int (f(x) - g(x))^2 \mu(dx) = 0.$$

Lema 1. *Sean f y g dos densidades de probabilidad con respecto a la medida μ. Entonces*

$$\int f(x) \ln f(x) \mu(dx) \geqslant \int f(x) \ln g(x) \mu(dx), \tag{1}$$

si estas dos integrales son finitas. El signo de igualdad sólo es posible en el caso de $f = g$ c.d. $[\mu]$.

Aquí se vino al acuerdo de que las integrales en (1) sobre el conjunto A, en el que $f(x) = 0$, equivalen a cero para cualquier $g(x)$.

Demostración. Es necesario demostrar que

$$\int f(x) \ln \frac{g(x)}{f(x)} \mu(dx) \leqslant 0.$$

Como $\ln(1 + x) \leqslant x$ para todos $x \geqslant -1$, y el signo de igualdad sólo es

posíble cuando $x = 0$, entonces

$$\ln \frac{g(x)}{f(x)} = \ln\left(1 + \left(\frac{g(x)}{f(x)} - 1\right)\right) \leqslant \frac{g(x)}{f(x)} - 1,$$

y el signo de igualdad aquí sólo es posible cuando $f(x) = g(x)$. Por eso

$$\int f(x) \ln \frac{g(x)}{f(x)}\, \mu(dx) \leqslant \int f(x)\left(\frac{g(x)}{f(x)} - 1\right)\mu(dx) =$$

$$= \int g(x)\mu(dx) - \int f(x)\mu(dx) = 0. \quad (2)$$

Si la relación $f = g$ c.d. $[\mu]$ no tiene lugar, es evidente que el signo de desigualdad en (2) será estricto. ◁

Examinemos ahora la familia $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ que satisface las condiciones (A_0), (A_μ) y la "distancia" $d(\mathbf{P}_\theta, \mathbf{Q})$ entre la distribución arbitraria $\mathbf{Q}$ y la distribución $\mathbf{P}_\theta \in \mathscr{P}$

$$d(\mathbf{P}_\theta, \mathbf{Q}) = -\int \ln f_\theta(x)\mathbf{Q}(dx). \quad (3)$$

Definamos la funcional $G(\mathbf{Q})$ como el valor de θ con el que se alcanza

$$\min_\theta d(\mathbf{P}_\theta, \mathbf{Q}) = d(\mathbf{P}_{G(\mathbf{Q})}, \mathbf{Q}).$$

Del lema 1 y la condición (A_0) se deduce que

$$-\int f_{\theta_0} \ln f_\theta \mu(dx) > -\int f_{\theta_0} \ln f_{\theta_0}\mu(dx),$$

$$d(\mathbf{P}_\theta, \mathbf{P}_{\theta_0}) > d(\mathbf{P}_{\theta_0}, \mathbf{P}_{\theta_0})$$

cuando $\theta \neq \theta_0$. Esto significa que

$$G(\mathbf{P}_{\theta_0}) = \theta_0. \quad (4)$$

Definición 1. Llámase *estimación de máxima verosimilitud* (*e.v.m.*) el valor de $\hat{\theta}^* = G(\mathbf{P}_n^*)$, o sea, el valor de θ con el que se alcanza

$$\max_\theta \int \ln f_\theta(x)\mathbf{P}_n^*(dx) = \max_\theta \frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(x_i). \quad (5)$$

En lo sucesivo, el símbolo $\hat{}$ sobre la designación de la estimación corresponderá siempre a la e.v.m.

De la definición y de (4) se deduce que la *e.v.m. es una estimación de sustitución.* Esta también puede ser considerada como la estimación con arreglo al valor mínimo de la distancia (3). Esta distancia se halla íntimamente ligada a la distancia de Kullback—Leibler entre las distribuciones, la cual desempeña un papel especial en la estadística matemática y será examinada más tarde.

En la definición 1, la familia $\{\mathbf{P}_\theta\}$ se supone tal que $\hat{\theta}^*$ sea una magnitud aleatoria *.

En vista de que el valor máximo de cierta función puede alcanzarse en varios puntos, la e.v.m., hablando en general, no es única. El ejemplo respectivo será expuesto un poco más tarde.

La denominación de dicha estimación está relacionada con la siguiente interpretación importante de la expresión

$$\sum_{i=1}^{n} \ln f_\theta(x_i) = \ln \prod_{i=1}^{n} f_\theta(x_i),$$

presente en (5). Para facilitar la exposición examinemos primero el caso discreto cuando μ es la medida de cálculo. Entonces $\prod_{i=1}^{n} f_\theta(x_i)$ es la probabilidad de que aparezca el resultado $X = (x_1, \ldots, x_n)$. Por lo tanto, elegimos, en calidad de $\hat{\theta}^*$, el valor del parámetro que *maximiza esta probabilidad* (pues las funciones $\varphi(\theta) > 0$ y $\ln \varphi(\theta)$ alcanzan los valores extremos en los mismos puntos).

Una interpretación análoga también tiene lugar en el caso general. En virtud de la independencia de x_i tenemos, para los conjuntos $B = B_1 \times \ldots \times B_n$, $B_i \in \mathfrak{B}_{\mathscr{X}}$

$$\mathbf{P}_\theta(X \in B) = \int_{B_1} f_\theta(x_1)\mu(dx_1) \ldots \int_{B_n} f_\theta(x_n)\mu(dx_n). \tag{6}$$

Recordemos que x_i, a distinción de los elementos de la muestra x_i, designan las variables aleatorias, y el vector $(x_1, \ldots, x_n)$ se designa a través de x. Supongamos que μ^n es el producto directo múltiplo de n de las medidas μ, así que $\mu^n(dx) = \prod_{i=1}^{n} \mu(dx_i)$. Entonces (6) significa que

$$\mathbf{P}_\theta(X \in B) = \int_B \prod_{i=1}^{n} f_\theta(x_i)\mu^n(dx)$$

y, por consiguiente, la función $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$ es la densidad de distribución del vector aleatorio X en $\mathscr{X}^n$ respecto a la medida μ^n,

$$\int_{\mathscr{X}^n} f_\theta(x)\mu^n(dx) = 1.$$

Ahora bien, $\prod_{i=1}^{n} f_\theta(x_i)\mu^n(dx)$ puede interpretarse (análogamente al caso

*) O sea, $\hat{\theta}^*$ realiza la aplicación medible de $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ en $(R^k, \mathfrak{B}^k)$.

discreto) como la probabilidad de que la muestra adquiera el valor del paralelepípedo formado por la intersección de las "franjas" $(x_i, x_i + dx_i)$, y la estimación de la máxima verosimilitud maximiza en θ esta probabilidad.

La función

$$f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i)$$

como función de θ se llama *función de verosimilitud,* y la función

$$L(X, \theta) = \ln f_\theta(X) = \sum_{i=1}^{n} l(x_i, \theta),$$

donde $l(x, \theta) = \ln f_\theta(x)$, se denomina *función logarítmica de verosimilitud.*

Esas mismas denominaciones de las funciones f y L también se utilizarán en el caso cuando como argumento, en vez de X, se halle el vector variable x. Ahora bien, la función de verosimilitud $f_\theta(x)$ es la función sobre $\mathscr{X}^n \times \Theta$ que, para cada $\theta \in \Theta$, constituye la densidad de la probabilidad respecto a la medida μ^n, así que la densidad $f_\theta x_1$) en $\mathscr{X}$ también es la función de verosimilitud para el caso $n = 1$.

Por otro lado, $f_\theta(X)$, por ejemplo, en el caso $\mathscr{X} = R$, puede considerarse como la función de verosimilitud de una muestra de volumen 1 en el caso multidimensional, cuando $\mathscr{X} = R^m = R^n$.

Cabe señalar que la e.v.m. no depende absolutamente de la elección de la medida μ, puesto que, al sustituir μ por cualquier medida equivalente μ_1, la función de verosimilitud $f_\theta(x)$ cambiará sólo en el factor $\frac{d\mu^n}{d\mu_1^n}(x)$ que no depende de θ.

Las propiedades asintóticas de la e.v.m. podrían haber sido investigadas en el mismo camino que utilizamos al estudiar las estimaciones por el método de momentos. Precisamente allí hemos aprovechado el hecho de que las estimaciones conforme al método de momentos son estadísticas de tipo I. Esto nos permitió determinar directamente su conciliabilidad fuerte y su normalidad asintótica. Al cumplirse ciertas condiciones para $f_\theta(x)$, las *e.v.m. serán estadísticas de tipo* II, y esto también permite (véanse los teoremas de los §§ 1.5, 1.8) determinar su conciliabilidad y su normalidad asintótica. No obstante, a nosotros nos será más cómodo estudiar directamente las propiedades de las e.v.d. (véanse los §§ 23—27), ya que esto permite realizar la investigación de un modo más económico y completo.

Hallemos las funciones de verosimilitud y las e.v.m. para algunas distribuciones expuestas en el § 2. En cuanto a las funciones de verosimilitud suaves, la manera más fácil de hallar su valor máximo consiste en igualar a cero las primeras derivadas.

Ejemplo 1. La distribución normal de Φ_{α,σ^2} en $\mathscr{X} = R$ tiene una densidad de

$$\varphi_{\alpha,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\alpha)^2}{2\sigma^2}}, \quad -\infty < \alpha < \infty, \ \sigma > 0.$$

Suponiendo, en este caso, que $\theta = (\alpha, \sigma^2)$, obtenemos

$$f_\theta(x) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \alpha)^2 \right\},$$

$$L(X, \theta) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \alpha)^2.$$

En vista de que ln es una función monótona, como ya hemos señalado, f y L alcanzan su valor máximo con los mismos valores de θ. Tenemos

$$\frac{\partial L}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \alpha),$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \alpha)^2.$$

Resolviendo, para el punto del valor máximo, el sistema de ecuaciones

$$\frac{\partial L}{\partial \alpha} = 0, \quad \frac{\partial L}{\partial \sigma} = 0,$$

obtenemos

$$\hat{\alpha}^* = \bar{x}, \quad (\hat{\sigma}^2)^* = S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Es fácil comprobar que en este punto realmente se alcanza el valor máximo de L.

Ejemplo 2. Examinemos la distribución Γ con densidad

$$\gamma_\alpha(x) = \frac{\alpha^\lambda}{G(\lambda)}\, x^{\lambda-1} e^{-\alpha x}, \ x \geqslant 0, \ \alpha > 0,$$

en el caso cuando se conoce el parámetro λ. Tenemos

$$L(X, \alpha) = \lambda n \ln \alpha - n \ln \Gamma(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i - \alpha \sum_{i=1}^{n} x_i,$$

$$\frac{\partial L}{\partial \alpha} = \frac{\lambda n}{\alpha} - \bar{x}n, \quad \hat{\alpha}^* = \lambda/\bar{x}.$$

Ejemplo 3. Tenemos la distribución binomial $\mathbf{B}_p$. Aquí, para $X \in \mathbf{B}_p$ tenemos que $\mathbf{P}(x_i = 1) = p$, $\mathbf{P}(x_i = 0) = 1 - p$,

$$f_p(X) = p^{\nu}(1 - p)^{n-\nu},$$

donde ν es el número de apariciones de 1 entre los elementos $x_1, \ldots, x_n$. Por lo tanto,

$$L(X, p) = \nu \ln p + (n - \nu) \ln(1 - p),$$

$$\frac{\partial L}{\partial p} = \frac{\nu}{p} - \frac{n - \nu}{1 - p}, \quad \hat{p}^* = \frac{\nu}{n}.$$

Proponemos al lector que procure, en forma de ejercicio, hallar las e.v.m. para todas las familias paramétricas expuestas en el § 2, y que las compare con las estimaciones según el método de momentos.

Ahora citaremos dos ejemplos de un tipo, algo diferente, cuando la función f_θ no es suave en θ y cuando no son vigentes los métodos de búsqueda de la e.v.m., relacionados con la derivación.

Ejemplo 4. Sea $X \in U_{\theta, 1+\theta}$ (distribución uniforme sobre $[\theta, 1 + \theta]$). Aquí

$$f_\theta(x) = \begin{cases} 1, & x \in [\theta, 1 + \theta], \\ 0, & x \notin [\theta, 1 + \theta], \end{cases}$$

$$f_\theta(X) = \begin{cases} 1, & \theta \leqslant x_{(1)} \leqslant x_{(n)} \leqslant 1 + \theta, \\ 0, & \text{de lo contrario,} \end{cases}$$

donde $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$ es la serie variacional. En este ejemplo, la estimación de verosimilitud máxima no es única. En efecto, $f_\theta(X) = 1$ (o sea, al valor máximo) para todos los valores de θ que satisfagan las relaciones $x_{(n)} - 1 \leqslant \theta \leqslant x_1$. Como $x_{(n)} - x_{(1)} \leqslant 1$, tales θ existen siempre. Podemos tomar, en particular, $\hat{\theta}^* = x_{(1)}$ o bien $\hat{\theta}^* = x_{(n)} - 1$.

Ejemplo 5. Sea $X \in \mathbf{U}_{0,\theta}$. Aquí

$$f_\theta(x) = \begin{cases} \theta^{-1}, & x \in [0, \theta], \\ 0, & x \notin [0, \theta], \end{cases}$$

$$f_\theta(X) = \begin{cases} \theta^{-n} \text{ si } x_i \in [0, \theta] \text{ para todos } i = 1, 2, \ldots, n. \\ 0, \quad \text{de lo contrario.} \end{cases}$$

Para obtener la forma de función $f_\theta(X)$ como función de θ, escribamos la condición $x_i \in [0, \theta)$, $i = 1, \ldots, n$, en la forma equivalente $\theta \geqslant \max x_i = x_{(n)}$. Así pues, $f_\theta(X) = 0$ cuando $\theta \in [0, x_{(n)})$, y $f_\theta(X) = \theta^{-n}$ cuando

$\theta \in (x_{(n)}, \infty)$. El gráfico de esta función se muestra en la fig. 1. Aquí, al igual que en el ejemplo precedente, la función f_θ es discontinua. El valor máximo de f_θ se alcanza en el punto $\hat{\theta}^* = x_{(n)}$.

Análogamente el lector puede hallar la e.v.m. para un parámetro bidimensional desconocido (α, β) cuando $X \in \mathbf{U}_{\alpha,\beta}$.

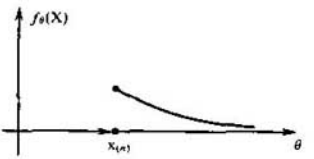

Fig. 1.

Si $f_\theta(x)$ es ilimitada y los puntos x_θ, en los que $f_\theta(x_\theta) = \infty$, dependen de θ, el método de verosimilitud máxima pierde en sumo grado su significado (aquí hemos venido al acuerdo de que $f_\theta(x_\theta) = \infty$ si $f_\theta(x) \to \infty$ cuando $x \downarrow x_\theta$ o cuando $x \uparrow x_\theta$). Esto se puede entender con más facilidad en el ejemplo del parámetro de desplazamiento cuando $f_\theta(x) = f(x - \theta)$, $f(x) > 0$, $f(0) = \infty$. Entonces $f_\theta(X) = \infty$ cuando $\theta = x_1, \ldots, \theta = x_n$ y, por consiguiente, $\hat{\theta}^*$ adquiere, por lo menos, n valores que coinciden con los elementos de la muestra. La esencia de tal efecto consiste en que en este caso los "saltos" de $f_\theta(X)$ no dan la posibilidad de juzgar acerca de la posición del máximo "verdadero" de $f_\theta(X)$, determinado por la influencia de toda la muestra (compárese esto con los §§ 24, 25). Para obtener tal parámetro sería necesario "amortiguar" de algún modo los saltos de $f_\theta(X)$.

Las estimaciones de verosimilitud máxima poseen la siguiente propiedad importante de invariancia con respecto a la sustitución del parámetro.

Teorema 1. *Supongamos que $\beta(\theta)$ es la función que realiza la aplicación biunívoca del conjunto Θ sobre el conjunto B. Entonces, si θ^* es la e.v.m. segun la muestra X del parámetro θ, en este caso $\beta^* = \beta(\hat{\theta}^*)$* será la e.v.m. según la muestra X del parámetro $\beta = \beta(\theta)$ *para la familia paramétrica $\{\mathbf{Q}_\beta = \mathbf{P}_{\theta(\beta)}\}_{\beta \in B}$, donde $\theta(\beta)$ es la función inversa a $\beta(\theta)$.*

Omitimos la **demostración** del teorema, debido a su evidencia.

Debemos señalar que ya hemos utilizado implícitamente el teorema 1 en el ejemplo 1, donde en busca de la e.v.m para σ^2 hemos hallado el valor máximo de L por σ y luego hemos tomado $(\tilde{\sigma}^2)^* = \hat{\sigma}^{*2}$.

Otro ejemplo de uso de este teorema es la determinación de la e.v.m. en el caso de $X \in \mathbf{L}_{\alpha,\sigma^2}$, o sea, en el caso cuando la distribución de x_i es lognormal: $\ln x_i \in \Phi_{\alpha,\sigma^2}$. Para tales x_i la media a y la varianza d^2 son iguales respectivamente (véase el § 2):

$$a = \exp\{\alpha + \sigma^2/2\}, \quad d^2 = a^2(e^{\sigma^2} - 1).$$

Si designemos por $\hat{a}^*$ y $(\hat{d}^2)^*$ las e.v.m. para a y d^2, en virtud de la propiedad de invariación obtenemos, para la función $(a, d^2) = \beta(\alpha, \sigma^2)$ (véase el ejemplo 1),

$$\hat{a}^* = \exp\left\{\overline{y} + \frac{S_Y^2}{2}\right\}, \quad (\hat{d}^2)^* = (\hat{a}^*)^2\,(e^{S_Y^2} - 1),$$

donde $Y = (y_1, \ldots, y_n)$, $y_i = \ln x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

El cálculo aproximado de las e.v.m. en situaciones más complicadas se realiza en el § 26.

Para resumir este párrafo haremos la observación siguiente. Ya hemos dicho que la e.v.m. es una estimación de sustitución. No obstante, dicha e.v.m. también puede considerarse, en ciertas condiciones, como estimación del método generalizado de momentos. En efecto, supongamos que la función $f_\theta(x)$ es derivable respecto a θ y que es legítima la derivación respecto a esta variable bajo el signo integral en la igualdad

$$\int f_\theta(x)\mu(dx) = 1.$$

Entonces

$$0 = \int \dot{f}_\theta(x)\mu(dx) = \int\limits_{\{f_\theta(x) \neq 0\}} \frac{\dot{f}_\theta(x)}{f_\theta(x)}\, f_\theta(x)\mu(dx) =$$

$$= \int\limits_{\{f_\theta(x) \neq 0\}} l'(x, \theta) f_\theta(x)\mu(dx) = \mathbf{M}_\theta l'(x_1, \theta).$$

Ahora bien, si en (4.6) ponemos $g(x, \theta) = l'(x, \theta)$, para la estimación por el método generalizado de momentos obtenemos la ecuación

$$\int l'(x, \theta)\mathbf{P}_n^*(dx) = \int l'(x, \theta)\mathbf{P}_\theta(dx) = 0$$

o bien, que es lo mismo,

$$L'(X, \theta) = 0.$$

Esta es la ecuación para la e.v.m.

§ 7. Acerca de la comparación de las estimaciones

Hemos visto que existen muchos enfoques naturales de obtención de las estimaciones. Cabe preguntar: ¿cómo comparar entre sí diferentes estimaciones y qué estimaciones deben preferirse a otras? Destaquemos dos enfoques de comparación de las estimaciones: estándar (medio cuadrático o típico) y asintótico.

El primero de ellos se basa en la comparación de las desviaciones estándar. El segundo enfoque es aplicable solamente a las muestras de gran volu-

men, puesto que se funda en la comparación de las "dispersiones" de las distribuciones para $(\theta^* - \theta)\sqrt{n}$ en caso de grandes n. Como base para tal comparación sirve generalmente la forma de distribuciones límite para $\theta^* - \theta)\sqrt{n}$ cuando $n \to \infty$ (si éstas existen). Los teoremas límite respectivos nos dan las condiciones en las que la distribución $(\theta^* - \theta)\sqrt{n}$ para grandes n puede ser aproximada con ayuda de las distribuciones límite mencionadas.

En este párrafo se supone que las estimaciones se comparan en caso de una distribución desconocida cualquiera de la muestra **P**, pero registrada.

1. Enfoque estándar. Caso unidimensional. Este enfoque se utiliza para examinar las estimaciones con arreglo a la muestra X de cualquier volumen registrado (no obligatoriamente grande). Consiste en la comparación de las desviaciones típicas $\mathbf{M}(\theta^* - \theta)^2$.

Regla 1. Con arreglo al enfoque estándar, consideraremos que la estimación θ_1^* es mejor que la θ_2^* si

$$\mathbf{M}(\theta_1^* - \theta)^2 < \mathbf{M}(\theta_2^* - \theta)^2.$$

Está ampliamente difundida la idea de que el error estándar es la característica numérica más conveniente de la exactitud de una estimación, aunque desde muchos puntos de vista esta circunstancia es discutible: pues se puede comparar, digamos, las magnitudes $\mathbf{M}|\theta^* - \theta|$ que también describen los valores medios de las desviaciones de θ^* de θ.

La ventaja indudable de las características $\mathbf{M}(\theta^* - \theta)^2$ consiste en el hecho de que $(\theta^* - \theta)^2$ es la función analítica de la diferencia $\theta^* - \theta$. Esto hace más cómodos muchos estudios y permite aproximar, como veremos más tarde, los valores de $\mathbf{M}f(\theta^* - \theta)$ para las funciones suaves f.

A la par con la desviación estándar para la descripción de las propiedades de las estimaciones también se utiliza la magnitud de desplazamiento.

Definición 1. Se llama *desplazamiento* de la estimación θ^* la magnitud

$$b = \mathbf{M}\theta^* - \theta.$$

La estimación θ^*, para la cual $b = 0$, se denomina *no desplazada*.

La desviación estándar está relacionada con el desplazamiento y la varianza de la estimación por medio de la igualdad

$$\mathbf{M}(\theta^* - \theta)^2 = \mathbf{D}\theta^* + b^2,$$

así que para las estimaciones no desplazadas, la desviación estándar coincide con la varianza.

El carácter de no desplazamiento propiamente dicho es, evidentemente, una propiedad deseable de las estimaciones, puesto que significa que en la sucesión dad de estimaciones, el valor medio de éstas coincidirá con

el valor verdadero del parámetro. Si falta dicha propiedad, la estimación se llama *desplazada.*

Ejemplo 1. Examinemos las tres estimaciones siguientes para el valor medio $\theta = \mathbf{M}x_1$ de la distribución $\mathbf{P}$:

$$\theta_1^* = \bar{x},\ \theta_2^* = \zeta^*,\ \theta_3^* = \frac{x_{(1)} + x_{(n)}}{2}, \tag{2}$$

donde ζ^* es la mediana muestral; $x_{(k)}$, $k = 1, \ldots, n$, los valores de la serie variacional, así que $\zeta^* = x_{((n+1)/2}$ si n es impar, y $\zeta^* = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$ si n es par (para $n = 1, 2$ todas las tres estimaciones coinciden). Todas las estimaciones son no desplazadas si la distribución $\mathbf{P}$, de la que ha sido extraida la muestra, es simétrica con respecto a $\theta(\mathbf{P}((-\infty, \theta - t)) = \mathbf{P}((\theta + t, \infty))$ para cualquier $t \geqslant 0$). Esto se deduce del hecho de que la distribución de todas las tres estimaciones también será simétrica respecto a θ. Para $\bar{x}$, la afirmación sobre el no desplazamiento de $\mathbf{M}\bar{x} = \theta$ es evidente incluso sin la suposición acerca de la simetría.

Calculemos las desviaciones estándar de las estimaciones (2). Para simplificar la exposición nos limitaremos al caso de $\mathbf{P} = \mathbf{U}_{0,1}$, $n = 3$, para el cual las estimaciones (2) pasarán a

$$\theta_1^* = \bar{x},\ \theta_2^* = x_{(2)},\ \theta_3^* = \frac{x_{(1)} + x_{(3)}}{2}.$$

Tenemos

$$\mathbf{D}x_1 = \int_0^1 (x - 1/2)^2 dx = 1/12,\ \mathbf{M}(\theta_1^* - \theta)^2 = \mathbf{D}\bar{x} = Dx_1/3 = 1/36.$$

Luego, en virtud de la definición de la mediana (n es impar) $\{\zeta^* < x\} = \{F_n^*(x) > 1/2\}$ y, por lo tanto

$$\mathbf{P}(\zeta^* < x) = \mathbf{P}(F_n^*(x) > 1/2) = \sum_{k=(n+1)/2}^{n} \mathbf{P}(nF_n^*(x) = k). \tag{3}$$

Para $n = 3$,

$$\mathbf{P}(F_3^*(x) = 1) = \mathbf{P}\Big(\bigcap_{i=1}^{3} \{x_i < x\}\Big) = F^3(x),$$

$$\mathbf{P}(3F_3^*(x) = 2) = 3F^2(x)(1 - F(x)).$$

La probabilidad $\mathbf{P}(\zeta^* \in (u, u + du))$ se compone de las probabilidades de sucesos que tienen la forma $\{x_1 \in (u, u + du)\}\{x_2 < u\}\{x_3 > u\}$. Como en total son posibles 6 de estas combinaciones, $\mathbf{P}(\zeta^* \in (u, u + du)) =$

$= 6f(u)F(u)(1 - F(u))du$ y, por consiguiente, ζ^* tiene una densidad igual a (esto también resulta de (3))

$$6f(u)F(u)(1 - F(u)),$$

donde $F(u) = \int_{-\infty}^{u} f(t)dt = \mathbf{P}(x_1 < u)$. En el caso de $\mathbf{P} = \mathbf{U}_{0,1}$ esta densidad será igual a $6x(1 - x)$ cuando $x \in [0, 1]$, así que

$$\mathbf{M}(\zeta^*)^2 = \int_0^1 6x^3(1 - x)dx = 6\left(\frac{1}{4} - \frac{1}{5}\right) = \frac{3}{10},$$

$$\mathbf{D}\zeta^* = \mathbf{M}(\zeta^*)^2 - (\mathbf{M}\zeta^*)^2 = \frac{3}{10} - \frac{1}{4} = \frac{1}{20}.$$

Nos queda hallar la varianza de la estimación

$$\theta_3^* = \frac{x_{(1)} + x_{(3)}}{2}.$$

Razonando análogamente a la precedente, no es difícil convencerse de que la probabilidad $\mathbf{P}(x_{(1)} \in (u, u + du), x_{(3)} \in (v, v + dv))$, cuando $u < v$, es igual a $6f(u)f(v)\ (F(v) - F(v))du\,dv$. Por eso para $\mathbf{P} = \mathbf{U}_{0,1}$

$$\mathbf{M}(\theta_3^*)^2 = \int_0^1 \int_0^v \left(\frac{u + v}{2}\right)^2 6(v - u)du\,dv.$$

El valor de esta integral es igual a 11/40 (el lector puede realizar los cálculos individualmente), por lo tanto,

$$\mathbf{D}\theta_3^* = \mathbf{M}(\theta_3^*)^2 - (\mathbf{M}\theta_3^*)^2 = \frac{11}{40} - \frac{1}{4} = \frac{1}{40}.$$

Así pues, la estimación θ_3^* resulta la mejor. Para otros valores de n y otras distribuciones $\mathbf{P}$, la situación puede ser otra. Veremos, por ejemplo, que cuando $\mathbf{P} = \Phi_{\alpha,\sigma^2}$, la mejor estimación para α será $\theta_1^* = \bar{x}$.

Ejemplo 2. *Estimaciones no desplazadas de la varianza.* Examinemos la estimación para la varianza

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

así como la estimación

$$S_1^2 = \frac{1}{n}\sum (x_i - \mathbf{M}x_i)^2 = \frac{1}{n}\sum x_i^2 + (\mathbf{M}x_1)^2 - 2\bar{x}\mathbf{M}x_1$$

(ambas según el principio de sustitución) en el caso cuando se conoce $\mathbf{M}x_1$.

La estimación S_1^2 no está, evidentemente, desplazada. Al mismo tiempo

$$S^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(x_i - \bar{x} \pm \mathbf{M}x_1)^2 =$$

$$= \frac{1}{n}\sum(x_i - \mathbf{M}x_1)^2 - (\bar{x} - \mathbf{M}x_1)^2 = S_1^2 - (\bar{x} - \mathbf{M}x_1)^2 < S_1^2.$$

Ahora bien, la estimación S^2 está desplazada,

$$\mathbf{M}S^2 = \mathbf{D}x_1 - \mathbf{D}\bar{x} = \left(1 - \frac{1}{n}\right)\mathbf{D}x_1.$$

Esta relación muestra que también podemos examinar, en caso de $\mathbf{M}x_1$ desconocida, la estimación de la varianza igual a

$$S_0^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2, \quad \mathbf{M}S_0^2 = \mathbf{D}x_1.$$

Pasemos ahora al enfoque asintótico del problema de comparación de las estimaciones. En este caso la regla para la preferencia de las estimaciones se elige unívocamente.

2. Enfoque asintótico. Caso unidimensional. Supongamos que se han dado dos estimaciones θ_1^* y θ_2^* tales que

$$\frac{(\theta_1^* - \theta)\sqrt{n}}{\sigma_1} \in \mathbf{Q}, \quad \frac{\theta_2^* - \theta)\sqrt{n}}{\sigma_2} \in \mathbf{Q}, \tag{4}$$

donde $\mathbf{Q}$ es cierta ley de distribución límite, la misma que para θ_1^* y θ_2^*, y $\sigma_2 > \sigma_1$. Entonces, para grandes valores de n, las distribuciones $(\theta_i^* - \theta)\sqrt{n}/\sigma_i$, $i = 1, 2$ serán próximas a $\mathbf{Q}$, e indudablemente que la "dispersión" de θ_2^* alrededor de θ será mayor que la "dispersión" de θ_1^* y debemos preferir θ_1^*.

Ahora bien, la esencia del enfoque asintótico consiste en la comparación de las distribuciones límites de las estimaciones.

Ya hemos visto y también nos convenceremos de ello ulteriormente, que muchas estimaciones aparecidas de un modo natural, icluyendo las óptimas (de lo cual hablaremos posteriormente), son asintóticamente normales, o sea, para ellas es válida (4) cuando $\mathbf{Q} = \Phi_{0,1}$. Esto nos permite enunciar la siguiente regla natural de comparación de las estimaciones a.n.

Supongamos que se dan dos estimaciones a.n. θ_1^* y θ_2^* con los coeficientes σ_1^2 y σ_2^2, respectivamente.

Regla 2. La estimación θ_1^* debe ser mejor que θ_2^* si $\sigma_1^2 < \sigma_2^2$.

En lo sucesivo, al utilizar estas y otras reglas, a la par con el término "mejor" también haremos uso, donde sea necesario, de las palabras "no peor", "peor", "no mejor" que corresponderán a los signos de desigualdad $\leqslant$, $>$, $\geqslant$ entre σ_1^2 y σ_2^2 (o bien entre $\mathbf{M}(\theta_1^* - \theta)^2$ y $\mathbf{M}(\theta_2^* - \theta)^2$ en (1)). Si

$\sigma_1^2 = \sigma_2^2$, diremos que estas estimaciones son asintóticamente equivalentes. El acuerdo propuesto es natural, y en las definiciones ulteriores no lo mencionaremos cada vez y sólo nos limitaremos a difinir la relación "mejor" o las relaciones semejantes a ésta.

Es preciso señalar que en la clase de estimaciones a.n., la minimalidad de la dispersión de θ^* quiere decir que la magnitud

$$\lim_{n\to\infty} \mathbf{P}(|\theta^* - \theta| < u/\sqrt{n})$$

será máxima para cada u. Esta circunstancia hace indiscutible la regla indicada para la comparación de las estimaciones a.n.

El enfoque asintótico, a pesar de su naturalidad, tiene una desventaja considerable: *sólo es aplicable a las estimaciones de gran volumen y únicamente en la clase de estimaciones a.n.*

Los dos enfoques señalados son, en cierto sentido, próximos uno a otro: en ambos casos el hecho se reduce a la comparación de las varianzas o de las magnitudes próximas a ellas. Por supuesto que la magnitud σ_1^2/n en (4), cuando $\mathbf{Q} = \Phi_{0,1}$, puede distinguirse considerablemente de $\mathbf{M}(\theta^* - \theta)^2$. Sin embargo, los ejemplos que ilustran este hecho (proponemos al lector que los construya él mismo) tienen, por lo común, carácter artificial.

La exposición ulterior de este capítulo está relacionada, en mucho, con la construcción de las estimaciones, óptimas para cada uno de los dos enfoques introducidos.

Ejemplo 3. Sea $X \in \Gamma_{\alpha,1}$. En el ejemplo 1 del § 4 hemos mostrado que ambas estimaciones

$$\alpha_1^* = (\overline{x})^{-1} \text{ y } \alpha_2^* = \left(\frac{1}{2n}\sum x_i^2\right)^{-1/2}$$

son estimaciones conforme al método de momentos. Además, α_1^* también es e.v.m. Luego hemos determinado que ambas estimaciones son asintóticamente naturales, con coeficientes α^2 y $\frac{5}{4}\alpha^2$, respectivamente, y por lo tanto, la estimación α_1^* es mejor que la α_2^* desde el punto de vista del enfoque asintótico. Ese mismo resultado, para $n \geqslant 2$, se obtiene cuando se trata del enfoque estándar.

Ahora citaremos un ejemplo que muestra que según las propiedades de la distribución, una misma estimación puede ser mejor o peor que alguna otra estimación registrada.

Ejemplo 4. Examinemos el problema de la estimación $\theta = \mathbf{M}x_1$ si se sabe que $X \in \mathbf{P}$ y que la distribución $\mathbf{P}$ es simétrica respecto al punto

θ (compárese con el ejemplo 1). En este caso la mediana de la distribución ζ coincide con θ. Examinemos también dos estimaciones para θ (ambas según el principio de sustitución): la media $\theta_1^* = \bar{x}$ y la mediana muestral $\theta_2^* = \zeta^*$. Supongamos, para precisar, que n es impar. Del corolario 2.2.1, cuando $k = (n+1)/2$, se deduce que si la función de distribución F es continuamente derivable en el punto $\theta = \zeta$, entonces

$$(\zeta^* - \zeta)\sqrt{n} \Rightarrow \frac{\xi}{2f(\theta)}, \quad \xi \in \Phi_{0,1}, \quad f(x) = F'(x).$$

Con otras palabras, en este caso ζ^* es la estimación a.n. con coeficiente $\sigma_2^2 = 1/(4f^2(\zeta))$.

Por otro lado, la estimación a.n. de $\bar{x}$ tiene por coeficiente $\sigma_1^2 = \mathbf{D}x_1$. Ahora bien, si

$$\int (x-\zeta)^2 dF(x) < \frac{1}{4f^2(\zeta)},$$

debemos preferir la estimación $\bar{x}$. Si el signo de desigualdad es inverso, entonces debemos preferir ζ^*. Cabe señalar que los números $\int (x-\zeta)^2 dF(x)$ y $f(\zeta)$ son características de distribución muy poco relacionadas entre sí.

Examinemos un importante caso particular, cuando estimamos el parámetro α por la muestra $X \in \Phi_{\alpha,\sigma^2}$. En este caso $f(\alpha) = f(\zeta) = \dfrac{1}{\sigma\sqrt{2\pi}}$, así que

$$\sigma_2^2 = \frac{\pi}{2}\sigma^2 > \sigma^2 = \sigma_1^2.$$

Esto significa que en esta situación, la estadística $\bar{x}$ es mejor que la ζ^*. Sin embargo, como hemos visto, no es difícil construir el ejemplo de la distribución para la cual será preferible la estadística ζ^*.

El ejemplo de la mediana también es muy aleccionador en otro sentido. El mismo muestra que la velocidad de disminución del grado de dispersión de $\zeta^* - \zeta$ puede ser cualquiera. Para cerciorarse de esto, basta con recurrir a la observaicón 2.2.1. En condiciones de dicha observación, como factor normalizador que asegura la convergencia de $\zeta^* - \zeta$ hacia la distribución límite sirve la magnitud $n^{1/(2\gamma)}$, donde γ es un número no negativo cualquiera (véase (2.12)). El factor $\sqrt{n}$ corresponde solamente a las distribuciones suaves.

Ahora presentaremos un experimento real con la muestra de volumen $n = 101$ de la población normal $\Phi_{0,1}$ y veremos *) cómo los valores de $\bar{x}$

*) La muestra X ha sido construida con ayuda de los números aleatorios tomados de las tablas [8] (se han utilizado los primeros 101 números en la página).

y ζ^* aproximan el 0 cuando $n = 11, 21, 51, 101$. Los datos obtenidos se ofrecen en la tabla siguiente:

n	11	21	51	101
$\bar{x}$	−0,283	−0,254	−0,148	−0,072
ζ^*	−0,291	−0,292	−0,078	−0,044

En este ejemplo, la estimación ζ^* para $n = 51, 101$ se comporta mejor, lo cual es resultado de la desviación aleatoria. Para convencerse de la ventaja de $\bar{x}$ sería necesario realizar muchos experimentos de este tipo.

Veamos ahora que aspecto tienen los dos enfoques (anteriormente enunciados) de la comparación de las estimaciones en el caso multidimensional, cuando θ es el vector $(\theta_1, \ldots, \theta_k)$.

3. Enfoques estándar y asintótico en el caso multidimensional. Como antes, utilizaremos el enfoque asintótico sólo en la clase de estimaciones a.n. En este caso el hecho se reduce por completo a la comparación de las distribuciones normales multidimensionales (distribuciones límites para $(\theta^* - \theta)\sqrt{n}$) que se describen totalmente por medio de la matriz de segundos momentos σ^2 (véase, por ejemplo, el teorema 3.2A).

Si se examina el enfoque estándar de la comparación de las distribuciones exactas de θ^*, también todo se reduce a la posibilidad de comparar dos distribuciones en R^k, basándose en el conocimiento de los momentos $(\theta^* - \theta)$ de segundo orden. Ahora bien, en ambos casos debemos saber comparar, según el "grado de dispersión", las matrices de los momentos de segundo orden.

Examinemos los métodos de comparación más naturales. Supongamos que $\mathbf{Q}_1$ y $\mathbf{Q}_2$ son dos distribuciones aleatorias en R^4. Designemos por ξ_1 y ξ_2 cualesquiera vectores aleatorios que poseen estas distribuciones: $\xi_i \in \mathbf{Q}_i$.

Definición 2. Diremos que la dispersión estándar de la distribución $\mathbf{Q}_1$ alrededor del punto $\alpha \in R^k$ no es mayor que la dispersión $\mathbf{Q}_2$ si para todo vector $a = (a_1, \ldots, a_k)$,

$$\mathbf{M}(\xi_1 - \alpha, a)^2 \leqslant \mathbf{M}(\xi_2 - \alpha, a)^2, \tag{5}$$

donde $(x, a) = \sum_{i=1}^{k} x_i a_i$ es el producto escalar.

Diremos que la dispersión para $\mathbf{Q}_1$ es menor que para $\mathbf{Q}_2$ si en (5) tiene lugar el signo de desigualdad estricta al menos para un a.

Si $\alpha = \mathbf{M}\xi_1 = \mathbf{M}\xi_2$, la igualdad (5) significa que por cualquier dirección de a la varianza de la distribución $\mathbf{Q}_1$ (o sea, la varianza de la proyección de ξ_1 sobre a) no supera la magnitud igual para $\mathbf{Q}_2$.

Si $d_l^2 = \|d_{ij}^{(l)}\|$ es la matriz de segundos momentos de $\mathbf{Q}_l$, $l = 1, 2$, entonces, abriendo paréntesis en (5) para $\alpha = 0$, obtenemos, para todos $a_1, \ldots, a_k$

$$\sum_{i,j=1}^{k} d_{ij}^{(1)} a_i a_j \leqslant \sum_{i,j=1}^{k} d_{ij}^{(2)} a_i a_j. \tag{6}$$

En el lenguaje de las matrices designaremos esta relación por

$$d_1^2 \leqslant d_2^2, \tag{7}$$

que significa la definición no negativa de la matriz $d_2^2 - d_1^2$.

Ahora bien, la dispersión estándar de $\mathbf{Q}_1$ alrededor del cero no supera tal dispersión para $\mathbf{Q}_2$ si y sólo si para las matrices de los momentos de segundo orden tienen lugar las desigualdades (6) y (7).

Las reglas de preferencia de las estimaciones en el caso multidimensional pueden enunciarse del modo siguiente.

Enfoque estándar: la estimación θ_1^* es mejor que la θ_2^* si la dispersión estándar de θ_1^* alrededor del punto θ es menor que la misma magnitud para θ_2^*.

Si d_l^2 es la matriz de segundos momentos $\theta_l^* - \theta$, la afirmación que dice que "la estimación θ_1^* es mejor que la θ_2^*" significa que $d_1^2 < d_2^2$.

Enfoque asintótico: la estimación θ_1^* es mejor que la θ_2^* si la dispersión estándar cerca del cero de la distribución límite para $(\theta_1^* - \theta)\sqrt{n}$ es menor que la misma magnitud para $(\theta_2^* - \theta)\sqrt{n}$.

En otros términos, si $(\theta_l^* - \theta)\sqrt{n} \in \Phi_{0,\sigma_l^2}$, entonces la afirmación de que "la estimación θ_1^* es mejor que la θ_2^*" quiere decir que $\sigma_1^2 < \sigma_2^2$.

Se puede mostrar que si θ_1^* y θ_2^* son dos estimaciones a.n. y θ_1^* es mejor que θ_2^*, entonces

$$\lim_{n\to\infty} \mathbf{P}((\theta_1^* - \theta)\sqrt{n} \in B) > \lim_{n\to\infty} \mathbf{P}((\theta_2^* - \theta)\sqrt{n} \in B) \tag{8}$$

para cualquier elipsoide central *)B.

Vemos que en ambos casos la comparación de las estimaciones se reduce al establecimiento de las igualdades para las matrices de los momentos de segundo orden. Cierta diferencia consiste en que en el primer caso los momentos no son obligatoriamente centrales.

Establezcamos ahora ciertas relaciones equivalentes a (6), (7).

---

*) Para abreviar convengamos en llamar *elipsoide* en R^k el dominio $\sum_{i,j=1}^{k} d_{ij}x_i x_j \leqslant c$, y *elipse*, la superficie $\sum_{i,j=1}^{k} d_{ij}x_i x_j = c$.

Pongamos

$$v(\theta^*) = \mathbf{M}(\theta^* - \theta)V(\theta^* - \theta)^{\mathrm{T}}$$

y designemos por $\mathfrak{B}_+$ el conjunto de todas las matrices $V = \|v_{ij}\|$ definidas no negativamente. Si $\|d_{ij}\|$ es la matriz de segundos momentos $\theta^* - \theta$, entonces, evidentemente, $v(\theta^*) = \sum v_{ij}d_{ij}$.

Lema 1. $d_1^2 \leqslant d_2^2$ *si y sólo si* $v(\theta_1^*) \leqslant v(\theta_2^*)$ *para cualesquier* $V \in \mathfrak{B}_+$.

Demostración. En una dirección la afirmación es evidente, ya que la matriz $V_a = \|a_i a_j\| \in \mathfrak{B}_+$, y para tal matriz,

$$v_a(\theta_i^*) = \mathbf{M}(\theta_i^* - \theta)V_a(\theta_i^* - \theta)^T = \sum a_i a_j d_{ij}^{(i)}$$

(véase (6)).

Para demostrar la afirmación en dirección contraria, señalemos que el orden parcial basado en las desigualdades (5) es invariante respecto a los ejes de revolución de las coordenadas. Es decir, si C es la matriz de transformación ortogonal y θ_1^* es mejor que θ_2^* para el parámetro θ, entonces θ_1^*C es mejor que θ_2^*C para el parámetro θC. Esto se deduce de las igualdades

$$(\theta_i^*C - \theta C,\ a) = ((\theta_i^* - \theta)C,\ a) = (\theta_i^* - \theta,\ aC^T)$$

y de la definición 2.

Supongamos ahora que $d_1^2 < d_2^2$, o sea,

$$\sum d_{ij}^{(1)}a_i a_j < \sum d_{ij}^{(2)}a_i a_j. \tag{9}$$

Esto quiere decir que $v(\theta_1^*) < v(\theta_2^*)$ para las matrices V que tienen la forma $V_a = \|a_i a_j\|$ y, por lo tanto, también para las matrices diagonales $V_{\mathrm{diag}} \in \mathfrak{B}_+$, puesto que estas últimas son representables en forma de la suma de k matrices que tienen la forma V_a. Supongamos ahora que V es una matriz arbitraria de $\mathfrak{B}_+$ y C es una transformación ortogonal tal que $C^TVC = V_{\mathrm{diag}}$. Entonces

$$v(\theta_1^*) = \mathbf{M}(\theta_1^* - \theta)V(\theta_1^* - \theta)^T = \mathbf{M}(\theta_1^* - \theta)CV_{\mathrm{diag}}C^T(\theta_1^* - \theta)^T.$$

De las dos observaciones hechas anteriormente y de (9) se deduce que el segundo miembro de esta igualdad es menor que

$$\mathbf{M}(\theta_2^* - \theta)CV_{\mathrm{diag}}C^T(\theta_2^* - \theta)^T = \mathbf{M}(\theta_2^* - \theta)V(\theta_2^* - \theta)^T = v(\theta_2^*). \quad \triangleleft$$

Existe también otro método de comparar la dispersión (véase [37]) que, sin embargo, supone que ambas distribuciones $\mathbf{Q}_1$ y $\mathbf{Q}_2$ no están degeneradas en R^k y tienen una media nula. En este caso las matrices de los segundos momentos centrales d_i^2 quedarán definidas positivamente y para ellas existen las inversas $A_i = (d_i^2)^{-1}$.

Supongamos que d^2 es la matriz de segundos momentos de la distribución $\mathbf{Q}$, y que $A = (d^2)^{-1}$.

Definición 3. Se llama *elipsoide de dispersión de la distribución* **Q** el elipsoide

$$tAt^T \leqslant k + 2$$

que entre todos los elipsoides se destaca unívocamente por su propiedad siguiente: si se examina la distribución uniforme **U** (o sea, la distribución en R^k con densidad constante dentro del elipsoide y con densidad nula fuera de éste), en este elipsoide, los primeros y segundos momentos de **Q** y de **U** coinciden (véase [25], p.333).

Lema 2. *Supongamos que las matrices* d_l^2, $l = 1, 2$, *no han sido degeneradas. La dispersión estándar de* $\mathbf{Q}_1$ *alrededor del cero no es mayor que la dispersión de* $\mathbf{Q}_2$ *si y sólo si el elipsoide de dispersión para* $\mathbf{Q}_1$ *se encuentra en el elipsoide para* $\mathbf{Q}_2$.

Demostración. Supongamos que la elipse $tA_1t^T = 1$ se encuentra en el interior de $tA_2t^T = 1$. Como es sabido, existe la transformación lineal no degenerada $t = uL$ que transfiere la elipse $1A_1t^T = 1$ a la esfera unitaria S_1, y la elipse $tA_2t^T = 1$, a la elipse S_2 con los ejes principales en dirección de los ejes de coordenadas. Esto quiere decir que $\tilde{A}_1 \equiv LA_1L^T = E$ (matriz unidad), $\tilde{A}_2 \equiv LA_2L^T = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_k^2)$, $0 < \lambda_j^2 \leqslant 1$, $j = 1, \ldots, k$. Como $\tilde{A}_1^{-1} = E$, $\tilde{A}_2^{-1} = \operatorname{diag}(\lambda_1^{-2}, \ldots, \lambda_k^{-2})$, la elipse $t\tilde{A}_2^{-1}t^T = 1$ será una inversión respecto a la esfera unitaria S_1 de la elipse S_2 y, por consiguiente, se encontrará en S_1. Como $\tilde{A}_2^{-1} = (L^T)^{-1}A_2L^{-1}$, entonces, efectuando la transformación "inversa" $u = tL^T$, obtenemos que la elipse $tA_1^{-1}t^T = td_1^2t = 1$ se halla fuera de $tA_2^{-1}t^T = td_2^2t^T = 1$. Evidentemente, la misma relación es válida para las elipses $td_1^2t^T = c$ y $td_2^2t^T = c$. Pero esto significa que la igualdad $td_1^2t^T = c$ conduce a $td_1^2t^T = c \leqslant td_2^2t^T$. La afirmación en dirección contraria se muestra exactamente de la misma manera. ◁

Ahora es importante señalar que, a distinción del caso unidimensional, la comparación de las dispersiones con ayuda de las matrices de segundos momentos sólo establece el orden parcial en el conjunto de todas las distribuciones. Por ejemplo, las matrices $d_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ y $d_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ no son ni mejor ni peor una que otra, ya que para el vector $a = (1, 0)$, (6) es válida, y para el vector $a = (0, 1)$, la desigualdad será inversa. Esto constituye una incomodidad considerable del orden introducido, aunque éste, como tal, no suscita dudas.

Podemos hacer muchas estimaciones (o muchas distribuciones) bien ordenadas, si comparamos, digamos, $\mathbf{M}|\theta^* - \theta|^2$, donde $|\cdot|$ es la norma euclídea en R^k, así que

$$\mathbf{M}|\theta^* - \theta^2| = \mathbf{M}\sum_{i=1}^{k}(\theta_i^* - \theta_i)^2. \tag{10}$$

Tal método de ordenación ya es discutible, puesto que en distintas circunstancias, la precisión en diversas direcciones puede apreciarse de modo diferente. Para considerar de algún modo esta circunstancia, se puede, en calidad de generalización, tener en cuenta la medida de exactitud

$$v(\theta^*) = \mathbf{M}(\theta^* - \theta)V(\theta^* - \theta)^T,$$

donde V es la matriz definida no negativamente (el caso (10) corresponde a $V = E$).

Del lema 1 se deduce que si la dispersión de θ_1^* alrededor de θ es menor que la dispersión de θ_2^*, entonces $v(\theta_1^*) < v(\theta_2^*)$. El caso inverso, hablando en general, es incorrecto: el cumplimiento de la desigualdad $v(\theta_1^*) < v(\theta_2^*)$ para una matriz cualquiera V (el orden completo propuesto más arriba se basa en una matriz registrada) no significa aún que la dispersión de θ_1^* alrededor de θ es menor que la dispersión de θ_2^*.

Pasemos ahora a examinar un importante caso *paramétrico*, cuando se estiman los parámetros desconocidos de las distribuciones de familias paramétricas.

§ 8. Comparación de las estimaciones en el caso paramétrico. Estimaciones eficientes

En el párrafo precedente hemos destacado dos enfoques (estándar y asintótico) de la comparación de la calidad de las estimaciones. Introduzcamos ahora algunos conceptos relacionados con estos enfoques en el caso paramétrico, cuando la distribución de la muestra X pertenece a cierta familia $\mathscr{P} = \{\mathbf{P}_\theta\}$. Al igual que antes, con los símbolos $\mathbf{M}_\theta$ y D_θ designamos la esperanza matemática y la varianza de la distribución $\mathbf{P}_\theta$.

1. Caso unidimensional. Recordemos que de acuerdo con el enfoque estándar debemos decir qu θ_1^* es mejor que θ_2^* si

$$d_1^2(\theta) = \mathbf{M}_\theta(\theta_1^* - \theta)^2 < \mathbf{M}_\theta(\theta_2^* - \theta)^2 = d_2^2(\theta). \tag{1}$$

Pero en el caso paramétrico, $d_l^2(\theta)$, $l = 1, 2$, son las funciones de θ y debemos decir "θ_1^* es mejor que θ_2^* *en el punto* θ" si $d_1(\theta) < d_2(\theta)$.

Análogamente sucede al utilizar el enfoque asintótico cuando se comparan las estimaciones a.n. para grandes volúmenes de la muestra n, confrontando sus distribuciones límites. La estimación θ_1^* se considera mejor que la θ_2^* en el punto θ, si en las relaciones

$$(\theta_l^* - \theta)\sqrt{n} \Subset \Phi_{0,\,\sigma_l^2(\theta)}, \quad l = 1, 2, \tag{2}$$

es justa $\sigma_1(\theta) < \sigma_2(\theta)$*).

*) Ya hemos señalado que en la amplia clase de casos $d_l^2(\theta) = n^{-1}\sigma_l^2 + o(n^{-1})$. Sin embargo, esto no se deduce de las definiciones de los números $d_l^2(\theta)$ y $\sigma_l^2(\theta)$.

Ahora bien, en ambos casos el problema de comparación de las estimaciones conduce al asunto de *comparación de las funciones,* digamos, $d_l(\theta)$, $\theta = \Theta$. Este conjunto no está ordenado, y en la clase de todas las estimaciones es posible introducir un orden parcial del modo siguiente.

Regla 1. La estimación θ_1^* es mejor que la θ_2^* si $d_1(\theta) \leqslant d_2(\theta)$ (o, respectivamente, $\sigma_1(\theta) \leqslant \sigma_2(\theta)$) para todos $\theta \in \Theta$ y al menos para un θ se cumple la desigualdad estricta $d_1(\theta) < d_2(\theta)$.

Si la estimación θ^* es tal que para ella existe la estimación θ_1^* que es mejor que θ^*, en estos casos se dice que θ^* es una *estimación inadmisible.*

Expongamos primeramente el enfoque estándar en el caso unidimensional y examinemos las posibilidades aquí existentes de comparar las estimaciones. Conviene señalar, ante todo, que desde el punto de vista de la definición citada no existe, hablando en general, la mejor estimación. O sea, no existe una estimación θ^* tal que para toda otra estimación θ_1^* sea válida la desigualdad $d(\theta) \leqslant d_1(\theta)$, donde $d_1(\theta)$ está definida en (1), y $d(\theta)$ corresponde a θ^*.

En efecto, si se toma la estimación $\theta_1^* = \theta_1 = \text{const} \in \Theta$, entonces $d_1^2(\theta) = \mathbf{M}_\theta(\theta_1^* - \theta)^2 = 0$ cuando $\theta = \theta_1$ y para la mejor estimación θ^* (si tal estimación existiera) se cumplirá $d^2(\theta_1) = \mathbf{M}_{\theta_1}(\theta^* - \theta_1)^2 = 0$. Como θ_1 es arbitrario, $d^2(\theta) \equiv 0$. Pero esto es posible únicamente en el caso "degenerado", cuando las observaciones determinan unívocamente el valor del parámetro θ. Por ejemplo, cuando $X \in \mathbf{I}_\theta$ o bien $X \in \mathbf{U}_{\theta, \theta+1}$ y $\Theta = \{1, 2, \ldots\}$.

Ahora bien, la envolvente inferior de todas las funciones $d^2(\theta)$ es igual a cero, pero en el caso "no degenerado" esta función no se realiza para ninguna función θ^*.

El problema puede ser más interesante si se buscan las mejores estimaciones θ^* en unas u otras subclases de estimaciones que se eligen de un modo suficientemente racional. Uno de los métodos posibles de destacar tales subclases consiste en registrar el desplazamiento $b(\theta)$.

Definición 1. La estimación $\theta_0^* \in K$ se denomina *eficiente en la clase* K si para cualquier otra estimación $\theta^* \in k$ $\mathbf{M}_\theta(\theta_0^* - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta)^2$ cuando todos $\theta \in \Theta$.

La clase K_0 de las estimaciones no desplazadas desempeña un papel especial, o sea, la clase de las estimaciones para las cuales $b(\theta) \equiv 0$.

Las estimaciones eficientes en la clase $K_0 = \{\theta^* : \mathbf{M}_\theta\theta^* = \theta\}$ de estimaciones no desplazadas se llaman simplemente *eficientes.* De suerte que las estimaciones eficientes no son sino estimaciones no desplazadas con varianza mínima.

Como ya hemos señalado, la propiedad de carácter no desplazado es, como tal, indudablemente deseable, ya que significa la falta del error sistemático al utilizar la estimación.

La cuestión acerca de la existencia de las estimaciones con el desplaza-

miento dado $b(\theta)$ (en particular, de las estimaciones no desplazadas) se reduce a la resolubilidad de la ecuación integral con respecto a $g(x)$:

$$\int g(x)\mathbf{P}_\theta(X \in dx) = \theta + b(\theta), \tag{3}$$

donde $g(X) = \theta^*$; el primer miembro de esta ecuación es $\mathbf{M}_\theta\theta^*$.

Si está cumplida la condición (A_μ) y $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$ es la función de verosimilitud, la ecuación toma la forma

$$\int g(x)f_\theta(x)\mu^n(dx) = \theta + b(\theta). \tag{4}$$

Cabe señalar que la solución (4) para $b(\theta)$ dada no siempre existe ni mucho menos y, en particular, no para *todas las familas* $\{\mathbf{P}_\theta\}$ *existen las estimaciones no desplazadas del parámetro* θ. Examinemos, por ejemplo, el esquema de Bernoulli con un parámetro desconocido p (la probabilidad del caso es $\{x_1 = 1\}$) y supongamos que nos hace falta estimar el parámetro $\theta = \varphi(p)$, donde φ es una función dada. Entonces la ecuación (4) para la estimación no desplazada tiene la forma

$$\sum g(x)f_\theta(x) = \theta$$

o bien, que es lo mismo,

$$\sum_{k=0}^{n} G(k)p^k(1-p)^{n-k} = \varphi(p), \tag{5}$$

donde $G(k) = \sum_{x \in A_k} g(x)$ y A_k es el conjunto de puntos x cuyas k coordenadas son iguales a 1. Pero el primer miembro de (5) es el polinomio de p de grado n. Esto significa que la ecuación (5) sólo puede ser resuelta si $\varphi(p)$ es un polinomio de grado no mayor de n.

Examinemos ahora la clase K_b de estimaciones con desplazamiento registrado $b(\theta)$ y supongamos que existe una estimación que es eficiente en K_b.

Teorema 1. *La estimación eficiente en* K_b *es única con una exactitud de hasta los valores sobre el conjunto* $A \subset \mathscr{X}^n$ *para el cual* $\mathbf{P}_\theta(A) = 0$ *cuando todos* $\theta \in \Theta$.

Demostración. Sean θ_0^*, θ_1^* dos estimaciones eficientes en K_b. Designemos

$$D = \mathbf{D}_\theta\theta_l^*, \ \Delta_l = \theta_l^* - \theta, \ \theta^* = \frac{\theta_0^* + \theta_1^*}{2}, \ l = 0, 1.$$

Como

$$\left(\frac{\Delta_0 + \Delta_1}{2}\right)^2 + \left(\frac{\Delta_0 - \Delta_1}{2}\right)^2 = \frac{\Delta_0^2 + \Delta_1^2}{2}$$

$$\frac{\Delta_0 + \Delta_1}{2} = \theta^* - \theta, \ \Delta_0 - \Delta_1 = \theta_0^* - \theta_1^*, \tag{6}$$

entonces

$$\mathbf{M}_\theta(\theta^* - \theta)^2 + \frac{1}{4}\,\mathbf{M}_\theta(\theta_0^* - \theta_1^*)^2 = D + b^2(\theta). \tag{7}$$

Pero $\theta^* \in K_b$ y, por lo tanto, $\mathbf{M}_\theta(\theta^* - \theta)^2 \geqslant D + b^2(\theta)$. En este caso, de (7) se deduce que

$$\mathbf{M}_\theta(\theta_0^* - \theta_1^*)^2 \leqslant 0,$$

$\theta_1^* = \theta_0^*$ c.s.*). ◁

El análisis realizado del problema de comparación de las estimaciones se refería al enfoque estándar. A este último también se refiere, en realidad, lo siguiente

Definición 2. La estimación $\theta_1^* \in K$ se denomina *asintóticamente eficiente* (a.e.) en K si cuando $n \to \infty$, para toda otra estimación θ^* de K y para cada $\theta \in \Theta$,

$$\lim_{n\to\infty} \sup \frac{\mathbf{M}_\theta(\theta_1^* - \theta)^2}{\mathbf{M}_\theta(\theta^* - \theta)^2} \leqslant 1. \tag{8}$$

Pasemos ahora al enfoque *asintótico* con el cual la definición 2 también está relacionada estrechamente. Aquí, como antes, el problema consiste en la comparación de las funciones $\sigma(\theta)$ que caracterizan la distribución normal límite, pero la cuestión en general se simplifica un poco. Esto se debe, ante todo, a que la comparación se realiza solamente en la clase de estimaciones a.n., que en lo sucesivo la designaremos por K_Φ. Podemos contraer un poco esta clase K_Φ sin empobrecerla considerablemente. Así pues, examinaremos la clase $K_{\Phi,2} \subset K_\Phi$ de las estimaciones a.n. θ^* que poseen la propiedad de que para ellas la convergencia

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2(\theta)}$$

ocurre junto con los dos primeros momentos:

$$\mathbf{M}_\theta(\theta^* - \theta)\sqrt{n} \to 0,\ \mathbf{M}_\theta(\theta^* - \theta)^2 n \to \sigma^2(\theta). \tag{9}$$

Señalemos que la primera de estas dos relaciones se obtiene fácilmente de

---

*) Es válida la siguiente afirmación que generaliza, en cierto sentido, el teorema 1. *Si θ_0^* es eficiente en k_b y la estimación θ^* es arbitraria en k_b, de modo que $h = D_\theta\theta_0^*/D_\theta\theta^* \leqslant 1$, entonces el coeficiente de correlación $\varrho(\theta_0^*, \theta^*)$ entre las estimaciones θ_0^* y θ^* es igual a*

$$\varrho(\theta_0^*, \theta^*) = \sqrt{h}.$$

El lector puede realizar individualmente la demostración, después de convencerse de que cuando $\varrho(\theta_0^*, \theta^*) \neq \sqrt{h}$ y al elegir correspondiente α, la estimación

$$\theta_1^* = (1 - \alpha)\theta_0^* + \alpha\theta^* \in K_b$$

satisfará la desigualdad $D_\theta\theta_1^* < D_\theta\theta_0^*$ que contradice la eficacia de θ_0^*.

la segunda con ayuda del teorema de continuidad para los momentos (§ 1.5).

La contracción de K_Φ hasta la clase $K_{\Phi,2}$ empobrece poco la primera de estas clases por dos causas. En primer lugar, las estimaciones a.n. en las que (9) no se cumple, prácticamente no existen (hemos señalado que para esto son necesarias, por regla general, construcciones artificiales). En segundo lugar, para $\theta^* \in K$, conforme al lema de Fatou,

$$\lim_{n\to\infty} \inf \mathbf{M}_\theta n(\theta^* - \theta)^2 \geqslant \sigma^2(\theta)$$

(se trata de las integrales de las funciones no negativas), así que $\mathbf{M}_\theta n(\theta^* - \theta)^2$, para grandes valores de n puede distinguirse de $\sigma^2(\theta)$ únicamente hacia el lado de los valores más grandes. Pero es poco probable que las estimaciones con tales propiedades puedan competir con las estimaciones para las cuales (9) ha sido cumplida.

Ahora bien, cuando se trata del enfoque asintótico, en calidad de clase de estimaciones a.n., en la cual se realiza la comparación, podemos considerar la clase $K_{\Phi,2}$. Esta será más cómoda para nosotros.

Sea K cierta clase de estimaciones, tal que $K \subset K_{\Phi,2}$. Entonces la siguiente definición será equivalente a la definición 2.

Definición 3. La estimación $\theta_1^* \in K$ se llama *asintóticamente eficiente en K, si para cualquier otra estimación* $\theta^* \in K$

$$\sigma_1^2(\theta) \leqslant \sigma^2(\theta) \tag{10}$$

cuando todos $\theta \in \Theta$, donde $\sigma^2(\theta)$ y $\sigma_1^2(\theta)$ son los coeficientes de dispersión de θ^* y θ_1^*, respectivamente.

La equivalencia de las definiciones se deduce del hecho de que para $\theta^* \in K_{\Phi,2}$

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \frac{\sigma^2(\theta)}{n}\,(1 + r_n(\theta)),\ r_n(\theta) \to 0 \text{ cuando } n \to \infty.$$

En este caso la relación (8), que significa que

$$\mathbf{M}_\theta(\theta_1^* - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta)^2(1 + r_n'(\theta)),\ r_n'(\theta) \to 0,$$

para cualquier $\theta^* \in K$ es, evidentemente, equivalente a la desigualdad (10). ◁

En el enfoque asintótico, cierta simplificación del problema de comparación (anteriormente recordada) consiste en que aquí comparamos tan sólo las varianzas de las leyes del límite. Aquí desaparece la importancia del desplazamiento $b(\theta)$ de las estimaciones, puesto que en la clase $K_{\Phi,2}$, en virtud de (9) se cumple la relación $b(\theta) = o(1/\sqrt{n})$ que significa "casi la falta de desplazamiento" de las estimaciones o la "despreciabilidad asintóti-

ca" del desplazamiento desde el punto de vista de las relaciones (2).

Análogamente al teorema 1 puede ser obtenido

El teorema 2. *Sea* $K \subset K_{\Phi,2}$. *Entonces, si* θ_1^* *y* θ_2^* *son dos estimaciones a.e. en K, tales que* $\frac{1}{2}(\theta_1^* + \theta_2^*) \in K$, *éstas coinciden asintóticamente, o sea,*

$$\sqrt{n}(\theta_1^* - \theta_2^*) \xrightarrow[P]{} 0, \ \mathbf{M}_\theta[\sqrt{n}(\theta_1^* - \theta_2^*)]^2 \to 0.$$

Demostración. Basta determinar la segunda relación, ya que la primera se deduce de ella. Sea

$$\mathbf{M}_{l,n} = \mathbf{M}_\theta n(\theta_l^* - \theta)^2, \ \Delta_l = \theta_l^* - \theta, \ \theta^* = \frac{\theta_1^* + \theta_2^*}{2}, \ l = 1, 2.$$

Entonces, en virtud de (6) obtenemos

$$\mathbf{M}_\theta n(\theta^* - \theta)^2 + \frac{1}{4}\mathbf{M}_\theta n(\theta_1^* - \theta_2^*)^2 = (M_{1,n} + M_{2,n})/2. \tag{11}$$

Pero $\theta^* \in K$ y, por consiguiente, después de pasar al límite, en la última igualdad obtenemos, en virtud de la eficacia asintótica de θ_l^*,

$$\lim_{n\to\infty} \mathbf{M}_\theta n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \ \triangleleft$$

Las consideraciones expuestas anteriormente contenían sólo una de las vías posibles de separar las estimaciones (en nuestro caso, las estimaciones eficientes) que, siguiendo varios razonamientos naturales, han de preferirse a otras. No obstante, son posibles, desde luego, también otros enfoques (recuérdese que teníamos que comparar los elementos no ordenados, o sea, las funciones $d(\theta)$ o $\sigma(\theta)$). Puesto que, hablando en general, no existen estimaciones con valores mínimos posibles de $d(\theta)$ para cada θ, entonces se pueden comparar, digamos, los valores medios $\int d(t)\,q(t)\,dt$, donde $q(t) \geqslant 0$, $\int q(t)\,dt = 1$, o los valores máximos $\max\limits_{\theta\in\Theta} d(\theta)$. Esto son los métodos de reglamentación de los conjuntos de todas las estimaciones.

Más tarde llamaremos *bayesiano* el primero de estos dos métodos, y *minimax*, el segundo. Las estimaciones óptimas bayesianas y minimax serán examinadas en el § 11, y las estimaciones eficientes, en los párrafos ulteriores.

El problema de elección de las estimaciones será examinado más detalladamente en el capítulo 5.

2. Caso multidimensional. Examinemos ahora el caso cuando θ y θ^* son vectores de R^k. Aquí, el problema de comparación de las estimaciones es más difícil. El hecho es que en el caso multidimensional teníamos que introducir un orden *parcial* ya para comparar las estimaciones cuando θ

ha sido *registrado*. Para comparar las estimaciones *en todo el conjunto* Θ, al igual que en el caso unidimensional, también es necesario introducir un orden parcial, pero ya "en otra dirección" puesto que la comparación se basa en la desviación estándar, que es una función de dos variables: θ y del vector a, sobre el cual se proyecta la desviación $\theta^* - \theta$).

Las mejores estimaciones en "ambas direcciones" constituyen precisamente el objeto de las definiciones siguientes.

Definición 4. La estimación θ_0^* es *eficiente en la clase* K si para cualquier estimación θ^* de K la dispersión estándar de θ^* alrededor de θ para todos $\theta \in \Theta$ no es menor que la dispersión de θ_0^*.

Esta definición es equivalente a la siguiente.

La estimación vectorial θ_0^* del parámetro θ es eficiente en K si para cualquier vector a la estimación $\alpha_0^* = (\theta_0^*, a)$ es la estimación eficiente del parámetro escalar $\alpha = (\theta, a)$ en la clase de estimaciones $\alpha^* = (\theta^*, a)$, $\theta^* \in K$, o sea, para todos $\theta \in \Theta$, $a \in R^k$, $\theta^* \in K$,

$$\mathbf{M}_\theta(\theta_0^* - \theta, a)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta, a)^2. \tag{12}$$

Como ya hemos visto, esta desigualdad se escribe de un modo equivalente en la forma $d_0^2(\theta) \leqslant d^2(\theta)$ o bien

$$\sum_{i,j} d_{ij}^{(0)}(\theta) a_i a_j \leqslant \sum_{i,j} d_{ij} a_i a_j$$

para todos $\theta \in \Theta$, $a \in R^k$, donde $d^2(\theta) = \|d_{ij}(\theta)\|$ y $d_0^2(\theta) = \|d_{ij}^{(0)}(\theta)\|$ son las matrices de segundos momentos $\theta^* - \theta$ y $\theta_0^* - \theta$, respectivamente.

Las estimaciones eficientes en la clase K_0 de las estimaciones no desplazadas se llaman simplemente *eficientes*.

En vista de que la definición (12) de la eficacia se construye a base de la utilización del caso unidimensional, estonces, mediante el teorema 1 no es difícil establecer que la estimación eficiente en la clase K_b de estimaciones, con un desplazamiento $b(\theta) = \mathbf{M}\theta^* - \theta$ registrado, es la única.

La definición de las estimaciones a.e. en el caso multidimensional es análoga a las definiciones 2 y 3.

Definición 5. La estimación vectorial θ_1^* del parámetro θ es *asintóticamente eficiente* en K si para cualquier vector a la estimación (θ_1^*, a) es la estimación a.e. del parámetro escalar $\alpha = (\theta, a)$ en la clase de estimaciones $\alpha^* = (\theta^*, a)$, $\theta^* \in K$.

En otros términos (véase el § 7), la dispersión estándar de la distribución límite $(\theta_1^* - \theta)\sqrt{n}$, para la estimación a.e. es mínima. Esto, a su vez, significa que para cualesquiera $\theta^* \in K$, $a \in R^k$, $\theta \in \Theta$ se cumple $\sigma_1^2(\theta) \leqslant \sigma^2(\theta)$, o bien

$$\sum_{i,j} \sigma_{ij}^{(1)}(\theta) a_i a_j \leqslant \sum_{i,j} \sigma_{ij}(\theta) a_i a_j,$$

donde $\sigma^2(\theta) = \|\sigma_{ij}(\theta)\|$, $\sigma_1^2(\theta) = \|\sigma_{ij}^{(1)}(\theta)\|$ son, respectivamente, las matrices de segundos momentos de las distribuciones límite $(\theta^* - \theta)\sqrt{n}$ y $(\theta_1^* - \theta)\sqrt{n}$.

Del párrafo precedente se puede sacar la conclusión de que el conjunto de estimaciones en el caso multidimensional, para θ registrado, puede ser ordenado si la calidad de la estimación se mide en cantidad (durante el enfoque estándar)

$$v(\theta^*) = \mathbf{M}_\theta(\theta^* - \theta)V(\theta^* - \theta)^T \equiv v(\theta^*, \theta), \tag{13}$$

donde V es la matriz definida no negativamente. La cantidad análoga relacionada con la matriz de segundos momentos de la distribución normal límite, también se puede examinar durante el enfoque asintótico en la clase $K_{\Phi,2}$.

Continuando el avance por este camino, es posible ordenar bien el conjunto de todas las estimaciones incluso en todo el conjunto Θ. A saber, se pueden comparar los valores medios

$$\int v(\theta^*, t)\, q(t)\, dt, \quad q(t) \geqslant 0, \quad \int q(t)\, dt = 1,$$

o los valores máximos $\max\limits_{t \in \Theta} v(\theta^*, t)$ de las cantidades $v(\theta^*, \theta)$ definidas en (13).

Si resulta que la estimación que es la mejor en tal enfoque, continúa siendo la mejor para cualquier matriz V definida no negativamente, esto significará, en virtud del lema 7.1, que esta estimación también será la mejor desde el punto de vista del orden parcial establecido en el § 7 (o sea, la desviación estándar mediada será la mínima en cualquier dirección).

Para construir las estimaciones óptimas en sentido de las definiciones examinadas en este párrafo, necesitaremos los conceptos y las propiedades de las esperanzas matemáticas condicionales y de las estadísticas suficientes.

§ 9. Esperanzas matemáticas condicionales

En este párrafo recordaremos la definición de las esperanzas matemáticas condicionales (e.m.c) y sus propiedades principales. Véase una exposición más completa en el suplemento III, así como en [11], [38[, [30], [61] y [84].

1. Definición de la e.m.c. Sean ξ y η dos variables aleatorias dadas en el espacio probabilístico $(\Omega, \mathfrak{F}, \mathbf{P})$.

La esperanza matemática condicional $\mathbf{M}(\xi/B)$ de la variable aleatoria ξ respecto al suceso B, $\mathbf{P}(B) > 0$, se define por la igualdad

$$\mathbf{M}(\xi/B) = \frac{\mathbf{M}(\xi;\, B)}{\mathbf{P}(B)}, \tag{1}$$

donde $\mathbf{M}(\xi;\ B) = \int_B \xi d\mathbf{P} = \mathbf{M}(\xi I_B)$, $I_B = I_B(\omega)$ es una variable aleatoria igual al indicador del conjunto B.

Admitamos que ξ y η son independientes, $B = \{\eta = x\}$ y $\mathbf{P}(B) > 0$. Entonces, para cualquier función medible $\varphi(x, y)$ conforme a (1),

$$\mathbf{M}[\varphi(\xi,\ \eta)/\eta = x] = \frac{\mathbf{M}\varphi(\xi,\ \eta)I_{\{\eta = x\}}}{\mathbf{P}(\eta = x)} = \frac{\mathbf{M}\varphi(\xi,\ x)I_{\{\eta = x\}}}{\mathbf{P}(\eta = x)} = \mathbf{M}\varphi(\xi,\ x) \quad (2)$$

La última igualdad es válida, ya que las variables aleatorias $\varphi(\xi, x)$ e $I_{\{\eta = x\}}$ como funciones de ξ y η, respectivamente, son independientes y, por consiguiente,

$$\mathbf{M}\varphi(\xi,\ x)I_{\{\eta = x\}} = \mathbf{M}\varphi(\xi,\ x)\mathbf{M}I_{\{\eta = x\}} = \mathbf{M}\varphi(\xi,\ x)\mathbf{P}(\eta = x).$$

Las relaciones (2) muestran que el concepto de e.m.c. también puede conservar su significado en el caso cuando la probabilidad de la condición es igual a 0: pues de por sí la igualdad

$$\mathbf{M}[\varphi(\xi,\ \eta)/\eta = x] = \mathbf{M}\varphi(\xi,\ x)$$

para ξ y η independientes se presenta natural, y con la suposición de $\mathbf{P}(\eta = x) > 0$ no está relacionada de ningún modo.

Supongamos que $\mathfrak{A}$ es la σ-álgebra de $\mathfrak{F}$. Vamos a definir ahora el concepto de e.m.c. de la variable aleatoria ξ con respecto a $\mathfrak{A}$ que designaremos por $\mathbf{M}(\xi/\mathfrak{A})$. Primero daremos la definición del caso "discreto", pero de modo que se generalice fácilmente.

Llamamos "discreto" el caso cuando la σ-álgebra de $\mathfrak{A}$ está formada (generada) no más que por una sucesión numerable de los sucesos disjuntos $A_1, A_2, \ldots$; $\cup A_i = \Omega$, $\mathbf{P}(A_i) > 0$. Este hecho se escribe en forma de $\mathfrak{A} = \sigma(A_1, A_2, \ldots)$ y significa que como elementos de $\mathfrak{A}$ sirven todas las uniones posibles de los conjuntos $A_1, A_2, \ldots$

Con ayuda de la variable aleatoria ξ y el sistema de sucesos $(A_1, A_2, \ldots)$ construiremos una nueva variable aleatoria $\hat{\xi} = \hat{\xi}(\omega)$ del modo siguiente:

$$\hat{\xi} = y_k \equiv \mathbf{M}(\xi/A_k) = \frac{\mathbf{M}(\xi;\ A_k)}{\mathbf{P}(A_k)} \text{ cuando } \omega \in A_k,\ k = 1, 2, \ldots$$

Con otras palabras,

$$\hat{\xi} = \sum_k \frac{\mathbf{M}(\xi;\ A_k)}{\mathbf{P}(A_k)} I_{A_k},$$

donde I_A es el indicador del conjunto A.

Definición 1. La variable aleatoria $\hat{\xi}$ se llama *e.m.c. de* ξ *con respecto a la* σ*-álgebra de* $\mathfrak{A}$ *y se designa por* $\mathbf{M}(\xi/\mathfrak{A})$.

Ahora bien, a distinción de las esperanzas matemáticas ordinarias, la e.m.c. $\mathbf{M}(\xi/\mathfrak{A})$ es una *variable aleatoria.* En nuestro caso esta variable es constante en los conjuntos A_k y equivale, en estos conjuntos, al promedio de ξ en A_k. Si ξ y $\mathfrak{A}$ son independientes (o sea, $\mathbf{P}(\xi \in B;\ A_k) = \mathbf{P}(\xi \in B)\mathbf{P}(A_k)$), entonces es evidente que $\mathbf{M}(\xi;\ A_k) = \mathbf{M}\xi\mathbf{P}(A_k)$ y $\hat{\xi} = \mathbf{M}\xi$.

Sin embargo, si $\mathfrak{A} = \mathfrak{F}$, entonces $\mathfrak{F}$ también es "discreta", ξ es constante en los conjuntos A_k y, por lo tanto, $\hat{\xi} = \xi$. Señalemos las dos propiedades principales siguientes de la e.m.c.:

1) $\hat{\xi}$ *es medible con respecto* a $\mathfrak{A}$.

2) *Para cualquier suceso* $A \in \mathfrak{A}$

$$\mathbf{M}(\hat{\xi};\ A) = \mathbf{M}(\xi;\ A).$$

La primera propiedad es evidente. La segunda se deduce del hecho de que todo suceso $A \in \mathfrak{A}$ es representable en la forma $A = \bigcup_k A_{jk}$ y, por consiguiente,

$$\mathbf{M}(\hat{\xi};\ A) = \sum_k \mathbf{M}(\hat{\xi};\ A_{j_k}) = \sum_k y_{j_k}\mathbf{P}(A_{j_k}) = \sum_k \mathbf{M}(\xi;\ A_{j_k}) = \mathbf{M}(\xi;\ A).$$

Esta propiedad es bastante clara: tras promediar la variable ξ respecto al conjunto A se obtiene el mismo resultado que al promediar la magnitud $\hat{\xi}$ ya promediada respecto a A_{j_k}.

Lema 1. *Las propiedades* 1) y 2) *definen unívocamente la e.m.c. y son equivalentes a la definición 1.*

Demostración. En una dirección la afirmación del lema ya está demostrada. Ahora supongamos que se han cumplido las condiciones 1 y 2. La mensurabilidad de $\hat{\xi}$ con respecto a $\mathfrak{A}$ quiere decir que $\hat{\xi}$ es constante en los conjuntos A_k. Designemos el valor de $\hat{\xi}$ sobre A_k a través de y_k. Como $A_k \in \mathfrak{A}$, de la propiedad 2 se deduce que

$$\mathbf{M}(\hat{\xi};\ A_k) = y_k\mathbf{P}(A_k) = \mathbf{M}(\xi;\ A_k)$$

y, por lo tanto, para $\omega \in A_k$

$$\hat{\xi} = y_k = \frac{\mathbf{M}(\xi;\ A_k)}{\mathbf{P}(A_k)}. \quad \triangleleft$$

Ahora podemos dar la definición general de la e.m.c.

Definición 2. Supongamos que ξ es una variable aleatoria en el espacio probabilístico $(\Omega, \mathfrak{F}, \mathbf{P})$ y que $\mathfrak{A} \subset \mathfrak{F}$ es la σ-subálgebra de $\mathfrak{F}$. Llámase *esperanza matemática condicional de* ξ *respecto* a $\mathfrak{A}$ la variable aleatoria $\hat{\xi}$ desig-

nada por $\mathbf{M}(\xi/\mathfrak{A})$, la cual posee las dos propiedades siguientes:

1) $\hat{\xi}$ *es medible respecto a* $\mathfrak{A}$.

2) *Para cualquier* $A \in \mathfrak{A}$ *es válida* $\mathbf{M}(\hat{\xi}; A) = \mathbf{M}(\xi; A)$.

En esta definición la variable aleatoria ξ puede ser tanto escalar como vectorial.

En seguida surgen las preguntas: ¿existe tal variable $\hat{\xi}$? y ¿es única ésta? Hemos visto que en el caso "discreto" la respuesta a estas preguntas es positiva. En el caso general es válido

Teorema 1. *Si* $\mathbf{M}|\xi|$ *es finita, entonces la función* $\hat{\xi} = \mathbf{M}(\xi/\mathfrak{A})$ *siempre existe en la definición 2 y es única con una exactitud de hasta los valores en el conjunto de probabilidad cero.*

Demostración. Primero supongamos que ξ es escalar, $\xi \geqslant 0$. Entonces la función del conjunto

$$\mathbf{Q}(A) = \int_A \xi d\mathbf{P} = \mathbf{M}(\xi; A), \quad A \in \mathfrak{A},$$

será la medida en $(\Omega, \mathfrak{A})$, que es absolutamente continua respecto a $\mathbf{P}$, puesto que $\mathbf{P}(A) = 0$ conduce a $\mathbf{Q}(A) = 0$. Por consiguiente, según el teorema de Radón—Nikodym ([11], Suplemento 3) existe la función $\mathfrak{A}$-medible $\hat{\xi} = \mathbf{M}(\xi/\mathfrak{A})$ única, con una exactitud de hasta los valores en el conjunto de medida cero, tal que

$$\mathbf{Q}(A) = \int_A \hat{\xi} d\mathbf{P}.$$

En el caso general pongamos $\xi = \xi^+ - \xi^-$, $\xi^+ = \text{máx}(0, \xi) \geqslant 0$, $\xi^- = \text{máx}(0, -\xi) \geqslant 0$,

$$\hat{\xi} = \hat{\xi}^+ - \hat{\xi}^-,$$

donde $\hat{\xi}^*$ es la e.m.c. para ξ^*. Esto demuestra la existencia de la e.m.c., ya que $\hat{\xi}$ satisfará las condiciónes 1) y 2) de la definición 2. De aquí también resulta la unicidad, ya que la suposición acerca de la no unicidad de $\hat{\xi}$ significará la no unicidad de $\hat{\xi}^+$ o de $\hat{\xi}^-$. La demostración para ξ vectoriales se reduce al caso unidimensional, ya que las propiedades 1) y 2) pertenecerán a las coordenadas de $\hat{\xi}$ cuya existencia y unicidad ya han sido demostradas. ◁

La esencia de la demostración citada es bastante clara: pues según la condición 2, para cualquier $A \in \mathfrak{A}$ se da $\mathbf{M}(\hat{\xi}; A) = \int_A \hat{\xi} d\mathbf{P}$, o sea, se dan los valores de las integrales de $\hat{\xi}$ de todos los conjuntos $A \in \mathfrak{A}$. Es evidente que esto debe definir unívocamente la función $\mathfrak{A}$-medible $\hat{\xi}$ con una exactitud de hasta los valores en el conjunto de medida 0.

El sentido de $\mathbf{M}(\xi/\mathfrak{A})$ queda el mismo y, en términos generales, constituye el promedio de ξ en los elementos "indivisibles" de $\mathfrak{A}$.

Si $\mathfrak{A} = \mathfrak{F}$, entonces, evidentemente, $\hat{\xi} = \xi$ satisface las propiedades 1) y 2) y, por lo tanto, $\mathbf{M}(\xi/\mathfrak{F}) = \xi$.

Definición 3. Supongamos que ξ y η son las variables aleatorias en $(\Omega, \mathfrak{F}, \mathbf{P})$ y que $\mathfrak{A} = \sigma(\eta)$ es la σ-álgebra engendrada por la variable aleatoria η. Entonces $\mathbf{M}(\xi/\mathfrak{A})$ también se llama esperanza matemática condicional de la variable ξ respecto a η.

A veces, para simplificar la exposición, en vez de $\mathbf{M}(\xi/\sigma(\eta))$ escribiremos $\mathbf{M}(\xi/\eta)$, lo cual no conduce a equivocaciones.

Como, por definición, $\mathbf{M}(\xi/\eta)$ es una variable $\sigma(\eta)$-medible aleatoria, esto significa (véase [11], p.65) que existe una función medible $g(x)$ para la cual

$$\mathbf{M}(\xi/\eta) = g(\eta). \tag{3}$$

Por analogía con el caso discreto, la magnitud $g(x)$ aquí puede ser interpretada como el resultado de la mediación de ξ en el conjunto $\{\eta = x\}$. Recordemos que en el caso discreto $g(x) = \mathbf{M}(\xi/\eta = x)$).

Definición 4. Si $\xi = I_C$ es el indicador del conjunto $C \in \mathfrak{F}$, entonces $\mathbf{M}(I_C/\mathfrak{A})$ se denominará *probabilidad condicional* $\mathbf{P}(C/\mathfrak{A})$ del suceso C respecto a $\mathfrak{A}$. Si $\mathfrak{A} = \sigma(\eta)$, entonces hablaremos de la probabilidad condicional $\mathbf{P}(C/\eta)$ del suceso C respecto a η.

Propiedades de la e.m.c.

1) *La e.m.c. posee propiedades de esperanzas matemáticas ordinarias* (véase [11], p.75), con la única diferencia de que las mismas se cumplen casi con seguridad (con probabilidad 1):

1a) $\mathbf{M}(c\xi/\mathfrak{A}) = c\mathbf{M}(\xi/\mathfrak{A})$ *si* $c = \text{const}$,

1b) $\mathbf{M}(\xi_1 + \xi_2/\mathfrak{A}) = \mathbf{M}(\xi_1/\mathfrak{A} + \mathbf{M}(\xi_2)/\mathfrak{A})$,

1c) *si* $\xi_1 \leqslant \xi_2$ *c.s.*, *entonces* $\mathbf{M}(\xi_1/\mathfrak{A}) \leqslant \mathbf{M}(\xi_2/\mathfrak{A})$.

2) *Es válida la desigualdad del tipo de Chébishev*: si ξ es real, $\xi \geqslant 0$, entonces para cualquier $x > 0$,

$$\mathbf{P}(\xi \geqslant x/\mathfrak{A}) \leqslant \frac{\mathbf{M}(\xi/\mathfrak{A})}{x}.$$

Lo mismo que las igualdades del punto 1, tal relación entre las e.m.c. se cumple casi con seguridad. Este mismo acuerdo será válido posteriormente para todas las relaciones entre las e.m.c.

3) *Si las* σ*-álgebras de* $\mathfrak{A}$ *y* $\sigma(\xi)$ *son independientes, entonces* $\mathbf{M}(\xi/\mathfrak{A}) = \mathbf{M}\xi$.

De aquí se deduce, en particular, que si ξ y η son independientes, entonces $\mathbf{M}(\xi/\eta) = \mathbf{M}\xi$. Si la σ-álgebra de $\mathfrak{A}$ es trivial, entonces, evidentemente, también obtenemos $\mathbf{M}(\xi/\mathfrak{A}) = \mathbf{M}\xi$.

4) *Para las e.m.c. son ciertos los teoremas de convergencia, válidos para las esperanzas matemáticas ordinarias,* por ejemplo, el teorema de convergencia monótona: si $\xi_n \uparrow \xi$, $\xi_n \geqslant 0$, entonces $\mathbf{M}(\xi_n/\mathfrak{A}) \uparrow \mathbf{M}(\xi/\mathfrak{A})$ c.s.

5) *Si η es escalar y medible respecto a $\mathfrak{A}$,* $\mathbf{M}|\xi| < \infty$, $\mathbf{M}|\xi\eta| < \infty$, entonces

$$\mathbf{M}(\eta\xi/\mathfrak{A}) = \eta\mathbf{M}(\xi/\mathfrak{A}).$$

Con otras palabras, las variables aleatorias $\mathfrak{A}$-medibles se comportan, respecto a la operación de e.m.c., como constantes (compararlo con la propiedad 1a).

6) *Para las e.m.c. quedan válidas todas las desigualdades principales para las esperanzas matemáticas ordinarias,* en particular, la desigualdad de Cauchy — Buniakovski

$$\mathbf{M}(|\xi_1\xi_2|/\mathfrak{A}) \leqslant [\mathbf{M}(\xi_1^2/\mathfrak{A})\mathbf{M}(\xi_2^2/\mathfrak{A})]^{1/2}$$

y la desigualdad de Jensen: si $\mathbf{M}|\xi| < \infty$, entonces para cualquier función $g(x)$ convexa hacia abajo,

$$g(\mathbf{M}(\xi/\mathfrak{A})) \leqslant \mathbf{M}(g(\xi)/\mathfrak{A}).$$

7) *Fórmula de la probabilidad completa* (propiedad 2 de la definición 2 cuando $A = \Omega$):

$$\mathbf{M}\xi = \mathbf{M}\mathbf{M}(\xi/\mathfrak{A}).$$

8) *Promediación sucesiva* (generalización de la propiedad 7)): si $\mathfrak{A} \subset \mathfrak{A}_1 \subset \mathfrak{F}$, entonces

$$\mathbf{M}(\xi/\mathfrak{A}) = \mathbf{M}(\mathbf{M}(\xi/\mathfrak{A}_1)/\mathfrak{A}).$$

En el Suplemento III se puede hallar la demostración de estas propiedades.

Es evidente que las propiedades 1), 3), — 5), 7) y 8) son válidas tanto para las variables aleatorias ξ escalares como para las vectoriales. Destacaremos especialmente la siguiente propiedad de las e.m.c.

9) Es sabido que la función $\varphi(\alpha) = \mathbf{M}(\xi - a)^2$ alcanza su valor mínimo cuando $a = \mathbf{M}\xi$ (véase, por ejemplo, [11]). Esa misma propiedad también es válida para la e.m.c.: *cuando $a(\omega) = \mathbf{M}(\xi/\mathfrak{A})$ se alcanza el valor mínimo $\mathbf{M}(\xi - a(\omega))^2$ entre todas las funciones $a(\omega)$ $\mathfrak{A}$-medibles.*

En efecto, $\mathbf{M}(\xi - a(\omega))^2 = \mathbf{M}\mathbf{M}((\xi - a(\omega))^2/\mathfrak{A}$, pero $a(\omega)$ se comporta como constante respecto a la operación $\mathbf{M}(\cdot/\mathfrak{A})$ (véase la propiedad 5)), así que

$$\mathbf{M}((\xi - a(\omega))^2/\mathfrak{A}) = \mathbf{M}((\xi - \mathbf{M}(\xi/\mathfrak{A}))^2/\mathfrak{A}) + \mathbf{M}((\mathbf{M}(\xi/\mathfrak{A}) - a(\omega))^2/\mathfrak{A})$$

y el valor mínimo de esta expresión se alcanza cuando $a(\omega) = \mathbf{M}(\xi/\mathfrak{A})$. Esta propiedad puede considerarse como definición de la e.m.c. equivalente a

la definición 2. Debido a ella, $\mathbf{M}(\xi/\mathfrak{A})$ puede interpretarse como la "proyección" de ξ sobre $\mathfrak{A}$.

La propiedad 9) admite la siguiente generalización para el caso multidimensional, cuando $\xi = (\xi_1, \ldots, \xi_s)$ es un vector aleatorio en R^s.

9A) *Sea $V = \|v_{ij}\|$ una matriz arbitraria, definida no negativamente y de dimensión $s \times s$, $a \in R^s$,*

$$\zeta(a) = (\xi - a)V(\xi - a)^T$$

(en particular, para $V = E$ obtenemos $\zeta(a) = |\xi - a|^2$). *Entonces, en la función $a(\omega) = \mathbf{M}(\xi/\mathfrak{A})$ se alcanza el valor mínimo $\min_{a \in A} \mathbf{M}\zeta(a)$ para la clase A de todas las funciones $\mathfrak{A}$-medibles.*

La demostración de este hecho transcurre igual que en el caso unidimensional. Designemos $\alpha = \mathbf{M}(\xi/\mathfrak{A})$. Entonces $\mathbf{M}\zeta(a) = \mathbf{M}\mathbf{M}(\zeta(a)/\mathfrak{A})$,

$$\begin{aligned}\mathbf{M}(\zeta(a)/\mathfrak{A}) = \mathbf{M}((\xi - a)V(\xi - a)^T/\mathfrak{A}) = \mathbf{M}((\xi - \alpha)V(\xi - \alpha)^T/\mathfrak{A}) + \\ + \mathbf{M}((\alpha - a)V(\xi - \alpha)^T/\mathfrak{A}) + \mathbf{M}((\xi - \alpha)V(\alpha - a)^T/\mathfrak{A}) + \\ + \mathbf{M}((\alpha - a)V(\alpha - a)^T/\mathfrak{A}).\end{aligned} \tag{4}$$

Como $\alpha - a$ es el vector $\mathfrak{A}$-medible, entonces, según la propiedad 5),

$$\mathbf{M}((\alpha - a)V(\xi - \alpha)^T/\mathfrak{A}) = (\alpha - a)V\mathbf{M}((\xi - \alpha)^T/\mathfrak{A}) = 0,$$
$$\mathbf{M}((\xi - \alpha)V(\alpha - a)^T/\mathfrak{A}) = [\mathbf{M}((\xi - \alpha)/\mathfrak{A})]V(\alpha - a)^T = 0.$$

En vista de que el último sumando en (4) no es negativo y equivale a cero cuando $a = \alpha$, la afirmación queda demostrada. ◁

§ 10. Distribuciones condicionales

A la par con las e.m.c., las distribuciones condicionales se pueden examinar respecto a las σ-subálgebras y respecto a las variables aleatorias. En este párrafo estudiaremos solamente las distribuciones condicionales respecto a las variables aleatorias.

Sean ξ y η dos variables aleatorias en $(\Omega, \mathfrak{F}, \mathbf{P})$ con valores en R^s y R^k, respectivamente, y sea $\mathfrak{B}^s$ la σ-álgebra de los conjuntos de Borel de R^s.

Definición 1. *La función $\mathbf{P}(B/y)$ de dos variables $y \in R^k$, $B \in \mathfrak{B}^s$ se llama distribución condicional de ξ, a condición de que $\eta = y$, si*

1) Para cada B $\mathbf{P}(B/\eta)$ es la probabilidad condicional $\mathbf{P}(\xi \in B/\eta)$ del suceso $\{\xi \in B\}$ respecto a η, o sea, $\mathbf{P}(B/y)$ es una función de Borel de y, tal que para cualquier $A \in \mathfrak{B}^k$,

$$\mathbf{M}(\mathbf{P}(B/\eta); \eta \in A) = \int_A \mathbf{P}(B/y)\mathbf{P}(\eta \in dy) = \mathbf{P}(\xi \in B, \eta \in A).$$

2) Para cada y, $\mathbf{P}(B/y)$ es la distribución de las probabilidades sobre B.

A veces escribiremos la función $\mathbf{P}(B/y)$ de una "forma más descodificada":

$$\mathbf{P}(B/y) = \mathbf{P}(\xi \in B/\eta = y).$$

Sabemos que para cada $B \in \mathfrak{B}^s$ existe una función de Borel $g_B(y)$ tal que $g_B(\eta) = \mathbf{P}(\xi \in B/\eta)$. Ahora bien, poniendo $\mathbf{P}(B/y) = g_B(y)$, satisfaremos la condición 1) de la definición. Sin embargo, en este caso la condición 2) no se deduce de ningún modo de las propiedades de la e.m.c. y de ninguna manera se ve obligada a ser cumplida: pues la probabilidad condicional $\mathbf{P}(\xi \in B/\eta)$ está definida para cada B, con una exactitud de hasta los valores en el conjunto N_B de medida cero (ya que existen muchas variantes de e.m.c.) y este conjunto puede ser propio para cada B. Por eso, si la unión

$N = \bigcup_{B \in \mathfrak{B}^s} N_B$ no tiene probabilidad nula, puede resultar que, por ejemplo, las igualdades

$$\mathbf{P}(\xi \in B_1 \cup B_2/\eta) = \mathbf{P}(\xi \in B_1/\eta) + \mathbf{P}(\xi \in B_2/\eta)$$

(aditividad de la probabilidad) a la vez para todos B_1, B_2 disjuntos de $\mathfrak{B}^s$ no se cumplen ni siquiera para un solo ω de N, o sea, en el ω-conjunto de N de una probabilidad positiva, la función $g_B(y)$ no será una distribución como la función B.

No obstante, en nuestro caso, *cuando ξ es una variable aleatoria con valores en R^s y con σ-álgebra de los conjuntos de Borel $\mathfrak{B}^s$, $g_B(\eta) = \mathbf{P}(\xi \in B/\eta)$, siempre se puede elegir de tal modo que $g_B(y)$ sea una distribución condicional* (véase [38], [30]).

Como era de esperar, las distribuciones condicionales poseen la propiedad natural consistente en que las e.m.c. se expresan en forma de integrales según las distribuciones condicionales.

Teorema 1. *Para toda función medible $g(x)$ que aplica R^s en R, tal que $\mathbf{M}|g(\xi)| < \infty$, es válida la igualdad*

$$\mathbf{M}(g(\xi)/\eta) = \int g(x)\mathbf{P}(dx/\eta). \tag{1}$$

Demostración. Es suficiente examinar el caso cuando $g(x) \geqslant 0$. Si $g(x) = I_A(x)$ es el indicador del conjunto A, entonces la fórmula (1) es evidentemente cierta, o sea, es cierta para cualquier función simple $g_n(x)$ (es decir, para una función que adopte un número finito de valores). Nos queda tomar la sucesión $g_n \uparrow g$ y utilizar la monotonía de ambos miembros en (1) y la propiedad 4) del § 9. ◁

En los problemas reales, para calcular las distribuciones condicionales, a menudo es posible valerse de la siguiente regla simple, que, para eviden-

ciar, podemos escribirla de la forma siguiente:

$$\mathbf{P}(\xi \in B/\eta = y) = \frac{\mathbf{P}(\xi \in B,\ \eta \in dy)}{\mathbf{P}(\eta \in dy} . \tag{2}$$

Por supuesto que ambas condiciones de la definición 1 serán satisfechas formalmente.

Si ξ y η tienen densidad de distribución, dicha igualdad adquirirá un sentido exacto.

Definición 2. Supongamos que la distribución condicional $\mathbf{P}(B/y)$, para cada y es absolutamente continua respecto a cierta medida μ en R^s:

$$\mathbf{P}(\xi \in B/\eta = y) = \int_B f(x/y)\mu(dx).$$

Entonces la densidad $f(x/y)$ se denomina *densidad condicional de* ξ (*respecto a la medida* μ), *a condición de que* $\eta = y$.

En otros términos, la función $f(x/y)$ medible conforme al par de variables x, y es la densidad condicional de ξ a condición de que $\eta = y$, si

1) Para cualesquiera conjuntos de Borel, $A \subset R^k$, $B \subset R^s$

$$\int_{y \in A} \int_{x \in B} f(x/y)\mu(dx)\mathbf{P}(\eta \in dy) = \mathbf{P}(\xi \in B,\ \eta \in A), \tag{3}$$

2) Para cada y la función $f(x/y)$ es la densidad de distribución de las probabilidades.

Del teorema 1 se deduce que si existe la densidad condicional, entonces

$$\mathbf{M}(g(\xi)/\eta) = \int g(x)f(x/\eta)\mu(dx).$$

Si suponemos adicionalmente que la distribución de η tiene una densidad $q(y)$ respecto a cierta medida λ en R^k, entonces (3) se puede escribir de la forma siguiente:

$$\int_{y \in A} \int_{x \in B} f(x/y)q(y)\mu(dx)\lambda(dy) = \mathbf{P}(\xi \in B,\ \eta \in A), \tag{4}$$

Examinemos ahora el producto directo de los espacios R^s y R^k y, a base de él, el producto directo de las medidas $\mu \times \lambda$ (si $C = B \times A$, $B \subset R^s$, $A \subset R^k$, entonces $\mu \times \lambda(C) = \mu(B)\lambda(A)$). En este espacio la relación (4) significa, evidentemente, que la distribución compatible de ξ y η en $R^s \times R^k$ tiene una densidad respecto a $\mu \times \lambda$, igual a

$$f(x,\ y) = f(x/y)q(y).$$

Pero también es válida la afirmación inversa.

Teorema 2. *Si la distribución compatible de ξ y η en $R^s \times R^k$ tiene una densidad $f(x, y)$ respecto a $\mu \times \lambda$, entonces la función*

$$f(x/y) = \frac{f(x, y)}{q(y)}, \quad \text{donde } q(y) = \int f(x, y)\mu(dx)$$

es la densidad condicional de ξ, a condición de que $\eta = y$, y la función $q(y)$ es la densidad de η respecto a la medida λ.

Demostración. La afirmación del teorema respecto a $q(y)$ es evidente, ya que $\int_A q(y)\lambda(dy) = \mathbf{P}(\eta \in A)$. Queda señalar que $f(x/y) = f(x, y)/q(y)$ satisface todas las condiciones en la definición 2 de la densidad condicional (la igualdad (4) equivalente a 3 está cumplida de un modo evidente). ◁

Observación 1. Las variables aleatorias ξ y η en el teorema 2 se pueden cambiar de lugar. Entonces obtendremos que, a la par con $f(x/y)$, existe la densidad condicional

$$q(y/x) = \frac{f(x, y)}{f(x)}, \ f(x) = \int f(x, y)\lambda(dy)$$

de la variable aleatoria η, a condición de que $\xi = x$. Este simple corolario del teorema 2 desempeñará un papel muy importante en la exposición posterior. Con arreglo a los problemas de la estadística, este corolario nos permitirá obtener, en el párrafo siguiente, la fórmula de Bayes que luego se utilizará con frecuencia a lo largo de todo este curso.

Ejemplo 1. Sea Φ_{α,σ^2} la distribución normal bidimensional de las variables ξ_1 y ξ_2, donde $\alpha = (\alpha_1, \alpha_2)$, $\alpha_1 = \mathbf{M}\xi_i$, $\sigma^2 = \|\sigma_{ij}\|$, $\sigma_{ij} = \mathbf{M}(\xi_i - \alpha_i)(\xi_j - \alpha_j)$, $i, j = 1, 2$. El determinante de la matriz de segundos momentos es igual a

$$|\sigma^2| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \varrho^2),$$

donde ϱ es el coeficiente de correlación entre ξ_1 y ξ_2. Ahora bien, si $|\varrho| \neq 1$, la matriz de segundos momentos no está degenerada y para ella existe la matriz inversa

$$A = (\sigma^2)^{-1} = \frac{1}{|\sigma^2|}\begin{Vmatrix} \sigma^{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{Vmatrix} = \frac{1}{1-\varrho^2}\begin{Vmatrix} \frac{1}{\sigma_{11}} & -\frac{\varrho}{\sqrt{\sigma_{11}\sigma_{22}}} \\ -\frac{\varrho}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{1}{\sigma_{22}} \end{Vmatrix}.$$

Por lo tanto, la densidad compatible de ξ_1 y ξ_2 (respecto a la medida de Lebesgue) es igual a (véase el § 2)

$$f(x,\ y) = \frac{1}{2\pi\sigma_{11}\sigma_{22}\sqrt{1-\varrho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-\alpha_1)^2}{\sigma_{11}} - \frac{2\varrho(x-\alpha_1)(y-\alpha_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(y-\alpha_2)^2}{\sigma_{22}}\right]\right\}.$$

Las densidades unidimensionales de ξ_1 y ξ_2 son, respectivamente, iguales a

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_{11}}}\, e^{-\frac{(x-\alpha_1)^2}{2\sigma_{11}}},\quad q(y) = \frac{1}{\sqrt{2\pi\sigma_{22}}}\, e^{-\frac{(y-\alpha_2)^2}{2\sigma_{22}}}.$$

Por eso la densidad condicional de ξ_1, a condición de que $\xi_2 = y$, es igual a

$$f(x/y) = \frac{f(x,\ y)}{q(y)} =$$

$$= \frac{1}{\sqrt{2\pi\sigma_{11}(1-\varrho^2)}} \exp\left\{-\frac{1}{2\sigma_{11}(1-\varrho^2)}\left(x-\alpha_1-\varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(y-\alpha_2)\right)^2\right\};$$

ésta es la densidad de la distribución normal, con un valor medio $\alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(y-\alpha_2)$ y la varianza $\sigma_{11}(1-\varrho^2)$. De aquí se deduce, en particular, que la e.m.c. de ξ_1 con respecto a ξ_2 es igual a

$$\mathbf{M}(\xi_1/\xi_2) = \alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(\xi_1 - \alpha_2).$$

La recta $x = \alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(y-\alpha_2)$ se llama línea de regresión de ξ_1 sobre ξ_2. La misma proporciona la mejor aproximación estándar de la variable ξ_1 para una $\xi_2 = y$ dada.

Ejemplo 2. Examinemos el problema consistente en calcular la densidad de la variable aleatoria $\xi = \varphi(\zeta,\ \eta)$, donde ζ y η son independientes. De la fórmula (3), cuando $A = R^k$, resulta que la densidad $f(x)$ de la distribución de ξ se expresa, mediante la densidad condicional $f(x/y)$, por la igualdad

$$f(x) = \int f(x/y)\mathbf{P}(\eta \in dy). \tag{5}$$

Con arreglo al problema sujeto a examen, por $f(x/y)$ es necesario entender la densidad de la variable aleatoria $\varphi(\zeta,\ y)$, puesto que $\mathbf{P}(\xi \in B/\eta = y) = \mathbf{P}(\varphi(\zeta,\ y) \in B)$.

La fórmula (5) suele ser muy útil al calcular las distribuciones de diferentes estadísticas. Por ejemplo, en el punto 6 del § 2 podríamos escribir directamente la fórmula (2.7) para la densidad de la distribución de Fisher sin deducirla de la forma de la función de distribución.

§ 11. Enfoques bayesiano y minimax de la estimación de los parámetros

La esencia del enfoque bayesiano consiste en que el parámetro desconocido θ se examina como *variable aleatoria* con cierta densidad (conocida o desconocida) de distribución $q(t)$, $t \in \Theta$, respecto a la medida λ, la cual, al igual que la medida μ en la condición (A_μ), será lo más a menudo la medida de Lebesgue o la medida de cálculo. La densidad $q(t)$ se llama densidad *a priori*, o sea, dada *antes* del experimento. El enfoque bayesiano supone que el parámetro desconocido θ se ha escogido aleatoriamente de la distribución de densidad $q(t)$.

Supongamos a continuación, que $f_t(x)$, $t \in \Theta$, $x \in \mathscr{X}^n$ es la función de verosimilitud introducida por nosotros en el § 6. Como ya hemos señalado, $f_1(x)$ es, para cada t, la densidad de distribución en $\mathscr{X}^n$. Por eso la función

$$f(x, t) = f_t(x)q(t)$$

es la densidad de cierta distribución en $\mathscr{X}^n \times \Theta$ respecto a la medida $\mu^n \times \lambda$ que puede interpretarse como la *densidad de distribución compatible de* X y θ. Con tal enfoque, en virtud del teorema 10.2, la función $f_t(x)$, $x \in \mathscr{X}^n$ *es la densidad condicional de* X a condición de que $\theta = t$:

$$f_t(x) = f(x/t), \quad \mathbf{M}_\theta g(X) = \mathbf{M}(g(X)/\theta).$$

En estos planteamientos, el aspecto formal del asunto exige que $f_t(x)$ sea una función medible en t y x. En lo sucesivo, por doquier donde esto sea necesario, supondremos que dicha propiedad tiene lugar.

Posteriormente, el parámetro, como variable aleatoria, siempre será designado por θ, mientras que para los valores registrados del parámetro utilizaremos las designaciones t, u, etc., así que

$$\mathbf{M}_t g(X) = \mathbf{M}(g(X)/\theta = t).$$

A la par con $f(x/t)$ podemos escribir la *densidad condicional* $q(t/x)$ *de la variable* θ *a condición de que* $X = x$:

$$q(t/x) = \frac{f_t(x)q(t)}{f(x)}, \quad f(x) = \int f_t(x)q(t)\lambda(dt). \tag{1}$$

Esta densidad define la llamada distribución *a posteriori* (o sea, *después del experimento*) de θ, que designaremos por $\mathbf{Q}_x$. La igualdad (1) se denomina *fórmula de Bayes* para la densidad de la distribución a posteriori. En lo sucesivo esta fórmula desempeñará un papel muy importante.

Con arreglo al caso bayesiano, la propiedad 9 de la e.m.c. significa lo siguiente: entre todas las funciones $\theta^* = \varphi(X)$ la mejor estimación para θ

(desde el punto de vista de minimización de $\mathbf{M}(\theta - \varphi(X))^2$) es la función

$$\theta_{\mathbf{Q}}^* = \mathbf{M}(\theta/X) = \int tq(t/X)\lambda(dt) = \int t\mathbf{Q}_X(dt). \tag{2}$$

Definición 1. La estimación $\theta_{\mathbf{Q}}^*$ definida por las fórmulas (2) y (1) se llama *bayesiana, correspondiente a la distribución a priori* $\mathbf{Q}$ *de densidad* $q(t)$.

Señalemos una vez más, que para la estimación bayesiana, la desviación estándar incondicional

$$\mathbf{M}(\theta^* - \theta)^2 = \mathbf{MM}((\theta^* - \theta)^2/\theta = \mathbf{MM}_\theta(\theta^* - \theta)^2 = \\ = \int \mathbf{M}_t(\theta^* - t)^2 q(t)\lambda(dt) \tag{3}$$

adopta el valor mínimo posible. La relación (3) muestra que la estimación bayesiana minimiza el valor medio (con una función ponderal dada $q(t)\lambda(dt)$) de la magnitud $\mathbf{M}_t(\theta^* - t)^2$.

Con otras palabras, si θ se escoge al azar, con densidad $q(t)$, *entonces la estimación bayesiana es la mejor desde el punto de vista del enfoque estándar.* La desviación estándar (3) de la estimación bayesiana puede representarse en la forma (véase (1)):

$$\mathbf{M}(\theta_{\mathbf{Q}}^* - \theta)^2 = \int \mathbf{M}_t(\theta_{\mathbf{Q}}^* - t)^2 q(t)\lambda(dt) = \\ = \iint (t - \theta_{\mathbf{Q}}^*)^2 f_t(x) q(t)\lambda(dt)\mu^n(dx) = \int \sigma_{\mathbf{Q}x}^2 f(x)\mu^n(dx) = \mathbf{M}\sigma_{\mathbf{Q}x}^2,$$

donde $\sigma_{\mathbf{Q}x}^2$ es la varianza de la distribución *a posteriori* $\mathbf{Q}_X$:

$$\sigma_{\mathbf{Q}x}^2 = \int (t - \theta_{\mathbf{Q}}^*)^2 q(t/X)\lambda(dt) = \int (t - \mathbf{M}(\theta/X))^2 \mathbf{Q}_X(dt). \tag{4}$$

El otro enfoque de la comparación de las estimaciones, que ya hemos señalado en el § 8, se basa en la comparación $\sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2$, donde $\Gamma \subset \Theta$ es un subconjunto dado de Θ (Γ coincide con Θ o es igual a aquella de sus partes respecto a la cual se ha logrado determinar que $\theta \in \Gamma$).

Definición 2. La estimación $\bar{\theta}^*$ se denomina *minimax* si para cualquier otra estimación θ^*

$$\sup_{t\in\Gamma} \mathbf{M}_t(\bar{\theta}^* - t)^2 \leqslant \sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2.$$

Con otras palabras, para la estimación minimax se alcanza

$$\inf_{\theta^*} \sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2 = \sup_{t\in\Gamma} \mathbf{M}_t(\bar{\theta}^* - t)^2. \tag{5}$$

Establezcamos ciertas relaciones útiles entre las estimaciones bayesianas y minimax.

Teorema 1. *Designemos por* $\theta_{\mathbf{Q}}^*$ *la estimación bayesiana para la distribución a priori* $\mathbf{Q}$ *de densidad* q. *Si existe la estimación* θ_1^* *y la distribución*

Q *tales que para todos t*

$$\mathbf{M}_t(\theta_1^* - t)^2 \leqslant \int \mathbf{M}_u(\theta_Q^* - u)^2 q(u)\lambda(du), \tag{6}$$

la estimación θ_1^* *es de tipo minimax.*

Demostración. Sea θ^* cualquier otra estimación. Entonces $\sup_t \mathbf{M}_t(\theta^* - t)^2 \geqslant \int \mathbf{M}_t(\theta^* - t)^2 q(t)\lambda(dt) \geqslant \int \mathbf{M}_t(\theta_Q^* - t)^2 q(t)\lambda(dt) \geqslant \mathbf{M}_t(\theta_1^* - t)^2$. ◁

Nótese que casi para todos t pertenecientes al portador $N_Q = \{t: q(t) > 0\}$ de la distribución **Q**, en la desigualdad (6) debe cumplirse indispensablemente la igualdad, ya que de lo contrario obtendríamos

$$\int \mathbf{M}_t(\theta_1^* - t)^2 q(t)\lambda(dt) < \int \mathbf{M}_t(\theta_Q^* - t)^2 q(t)\lambda(dt)$$

lo cual contradice la definición de la estimación bayesiana.

Esta observación nos permite enunciar el siguiente criterio del carácter minimax de la estimación, equivalente al teorema 1.

Teorema 2. *Si la estimación* θ^*

1) *es bayesiana para cierta distribución* **Q**,

2) $\mathbf{M}_t(\theta^* - t)^2 = c = \text{const}$ *para* $t \in N_Q$,

3) $\mathbf{M}_t(\theta^* - t)^2 \leqslant c$ *para los demás t, entonces* θ^* *es una estimación minimax.*

Si $\theta^* = \theta_Q^* = \overline{\theta}^*$ satisface este criterio, es evidente que

$$\sup_t \mathbf{M}_t(\overline{\theta}^* - t)^2 = \int \mathbf{M}_t(\theta^* - t)^2 q(t)\lambda(dt). \tag{7}$$

Ahora bien, la estimación minimax es una estimación bayesiana que "iguala" los errores $\mathbf{M}_t(\overline{\theta}^* - t)^2$ para diferentes t. Esto quiere decir que la distribución a priori $\overline{\mathbf{Q}}$, correspondiente a dicha estimación, obliga a ser igualmente atentos a todos los valores posibles de θ sin orientarse, como lo hacen las estimaciones bayesianas θ_Q^* correspondientes a otras distribuciones a priori $\mathbf{Q} \neq \overline{\mathbf{Q}}$, hacia ciertos valores destacados (más probables) de θ. En vista de que en el último caso utilizamos una información complementaria acerca de θ, es natural que para $\mathbf{Q} \neq \overline{\mathbf{Q}}$ las estimaciones θ_Q^* posean desviaciones estándar incondicionales de menores valores:

$$\int \mathbf{M}_t(\theta_Q^* - t)^2 \, \mathbf{Q}(dt) \leqslant \int \mathbf{M}_t(\theta_{\overline{Q}}^* - t)^2 \overline{\mathbf{Q}}(dt).$$

Por eso la distribución $\overline{\mathbf{Q}}$ en el teorema 2, la cual corresponde a la estimación minimax $\overline{\theta}^*$, a menudo se llama distribución *pésima*.

En vista de que tal distribución pésima $\overline{\mathbf{Q}}$ no siempre existe (eso suele suceder en los casos cuando Θ es un conjunto ilimitado), se puede proponer el siguiente criterio modificado para determinar la estimación minimax.

Teorema 3. *Si existe la estimación* θ_1^* *y la sucesión de distribuciones*

$\mathbf{Q}^{(k)}$ *con densidades* $q^{(k)}$ *tales que para todos* t

$$\mathbf{M}_t(\theta_1^* - t)^2 \leqslant \lim_{k\to\infty} \sup \int \mathbf{M}_t(\theta^*_{\mathbf{Q}}(k) - t)^2 q^{(k)}(t)\lambda(dt),$$

entonces la estimación θ_1^* *es minimax.*

La demostración de este teorema es igualmente simple. Para toda estimación θ^* es válida

$$\sup_t \mathbf{M}_t(\theta^* - t)^2 \geqslant \int \mathbf{M}_t(\theta^* - t)^2 q^{(k)}(t)\lambda(dt) \geqslant \int \mathbf{M}_t(\theta^*_{\mathbf{Q}^{(k)}} - t)^2 q^{(k)}(t)\lambda(dt).$$

De aquí se deduce que

$$\sup_t \mathbf{M}_t(\theta^* - t)^2 \geqslant \lim_{k\to\infty} \sup \int \mathbf{M}_t(\theta^*_{\mathbf{Q}^{(k)}} - t)^2 q^{(k)}(t)\lambda(dt) \geqslant \mathbf{M}_t(\theta_1^* - t)^2. \quad \triangleleft$$

Ejemplo 1. Sea $X \in \Phi_{\alpha,1}$. Determinemos qué representa la estimación bayesiana $\alpha^*_{\mathbf{Q}^{(k)}}$ del parámetro α con una distribución normal a priori $\mathbf{Q}^{(k)} = \Phi_{(0,k)}$. En este caso debemos poner $\lambda(dt) = dt$,

$$q^{(k)}(t) = \frac{1}{\sqrt{2\pi k}}\, e^{-\frac{t^2}{2k}}.$$

La distribución a posteriori $\mathbf{Q}_X^{(k)}$ tendrá una densidad $q^{(k)}(t/X)$ proporcional (como función de t) a $q^{(k)}(t)f_t(X)$ o bien, que es lo mismo, proporcional a

$$\exp\left\{-\frac{t^2}{2k} - \frac{1}{2}\sum(x_i - t)^2\right\}.$$

De la igualdad

$$-\frac{t^2}{2}\left(\frac{1}{k} + n\right) + \bar{x}nt = -\frac{1}{2}\left(\frac{1}{k} + n\right)\left(t - \frac{\bar{x}n}{\frac{1}{k} + n}\right)^2 + \frac{(\bar{x}n)^2}{2\left(\frac{1}{k} + n\right)}$$

se deduce que

$$\mathbf{Q}_X^{(k)} = \Phi_{\frac{\bar{x}nk}{1+nk},\ \frac{k}{1+nk}}.$$

Como la estimación bayesiana $\alpha^*_{\mathbf{Q}^{(k)}}$ del parámetro α es igual a la esperanza matemática de la distribución a posteriori, de aquí obtenemos

$$\alpha_{\mathbf{Q}^{(k)}} = \frac{\bar{x}nk}{1 + nk} = \frac{\bar{x}}{1 + \frac{1}{nk}}.$$

La varianza de la distribución a posteriori $\sigma^2_{\mathbf{Q}_X^{(k)}} = \frac{k}{1 + nk}$ no depende

de X. Por consiguiente, en virtud de (4) el error estándar de la estimación bayesiana es igual a

$$\frac{k}{1+nk} \to \frac{1}{n}$$

cuando $k \to \infty$. Por eso para la estimación $\alpha^* = \bar{x}$ tenemos

$$\mathbf{M}_t(\bar{x} - t)^2 = \frac{1}{n} = \lim_{k\to\infty} \int \mathbf{M}_t(\alpha^*_{Q^{(k)}} - t)^2 q^{(k)}(t)dt$$

y, por lo tanto, según el teorema 3, la estimación $\alpha^* = \bar{x}$ es minimax. La distribución "pésima" sería aquí la distribución uniforme en toda la recta (distribución "límite" para $\Phi_{o,k}$), si tal distribución existiera *).

En el ejemplo siguiente, el conjunto Θ es compacto y existe la distribución "pésima".

Ejemplo 2. Supongamos que $X \in B_p$, o sea, que x_j, $j = 1, \ldots, n$ adoptan los valores 1 y 0, respectivamente, con probabilidades p y $1 - p$, $p \in \Theta = [0, 1]$. Como sabemos, en este caso para la estimación $p^* = \bar{x}$ es válida

$$\mathbf{M}_p(\bar{x} - p)^2 = p(1-p)/n,$$

así que el criterio del teorema 2 no se ha cumplido. Examinemos la estimación

$$p^* = \frac{\bar{x} + \dfrac{1}{2\sqrt{n}}}{1 + \dfrac{1}{\sqrt{n}}}. \tag{8}$$

Para ella el error

$$\mathbf{M}_p(p^* - p)^2 = \left(1 + \frac{1}{\sqrt{n}}\right)^{-2} \mathbf{M}_p\left(\bar{x} - p + \frac{1}{2\sqrt{n}} - \frac{p}{\sqrt{n}}\right)^2 =$$

$$= \frac{n}{(1+\sqrt{n})^2}\left(\frac{p(1-p)}{n} + \frac{(1-2p)^2}{4n}\right) = \frac{1}{4(1+\sqrt{n})^2}$$

no depende de p. Si ahora nos convencemos de que la estimación (8) es bayesiana, determinaremos de este modo su carácter minimax. Examinemos la distribución a priori $\mathbf{Q} = \mathbf{B}_{N+1,N+1}$, donde $\mathbf{B}_{\lambda_1,\lambda_2}$ es la distribución

*) Es interesante anotar que la estimación $\alpha^* = \bar{x}$ deja de poseer la propiedad mencionada, si x es una muestra de una distribución normal multidimensional cuya dimensión constituye más de dos ($x_j \in R^k$, $\alpha \in R^k$, $k \geqslant 3$). Esto se expone más detalladamente en [48].

beta de densidad (véase el punto 8 del § 2)

$$\frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, t^{\lambda_1 - 1}(1 - t)^{\lambda_2 - 1}.$$

Entonces, como

$$f_t(X) = t^{\bar{x}n}(1 - t)^{n(1 - \bar{x})},\ q(t) = \frac{\Gamma(2N + 2)}{\Gamma^2(N + 1)}\, t^N(1 - t)^N,$$

la distribución a posteriori tendrá una densidad $q(t/X)$ que, como función de t, será proporcional a $f_t(X)q(t)$ o bien, que es lo mismo, será proporcional a

$$t^{N + \bar{x}n}(1 - t)^{N + (1 - \bar{x})n}.$$

Esto significa que la distribución a posteriori coincide con $\mathbf{B}_{N + \bar{x}n + 1, N + n(1 - \bar{x}) + 1}$. En vista de que el valor medio de la distribución $\mathbf{B}_{\lambda_1, \lambda_2}$ es igual a $\lambda_1/\lambda_1 + \lambda_2$) (véase el punto del § 2), la estimación bayesiana p_Q, correspondiente a $\mathbf{Q}$, será igual a

$$p_Q^* = \frac{N + \bar{x}n + 1}{2N + n + 2} = \frac{\bar{x} + (N + 1)/n}{1 + 2(N + 1)/n}.$$

Cuando $N + 1 = \sqrt{n}/2$, está estimación coincidirá con la estimación p^* definida en (8) y, en virtud del teorema 2, será minimax. La distribución $\mathbf{Q}$ será la peor (pésima), ya que se concentra o medida que crece n alrededor del "peor" valor del parámetro $p = 1/2$ con el que la varianza de la estimación $\bar{x}$, igual a $p(1 - p)/n = 1/(4n)$, será máxima. La propia estimación $\bar{x}$ no es minimax, ya que

$$\sup_p \frac{p(1 - p)}{n} = \frac{1}{4n} > \frac{1}{4(1 + \sqrt{n})^2}.$$

Al mismo tiempo es natural que para todos los valores de p que están fuera del entorno estrecho del punto $p = 1/2$, la estimación $\bar{x}$ será, sin embargo, mejor que p_Q^*, y esto tendrá lugar para todos los valores p para los cuales

$$p(1 - p) < \frac{1}{4(1 + 1/\sqrt{n})^2}.$$

En el caso general la determinación de las expresiones exactas (funciones explícitas de X) para las estimaciones bayesianas y minimax no es siempre posible. Por eso es natural utilizar también el enfoque asintótico.

Antes de introducir las definiciones correspondientes, debemos recordar que las estimaciones bayesianas y minimas θ_Q^* y $\bar{\theta}^*$ han sido definidas por

las desigualdades

$$\mathbf{M}(\theta_Q^* - \theta)^2 - \mathbf{M}(\theta^* - \theta)^2 \leqslant 0,$$
$$\sup_{t\in\Gamma} \mathbf{M}_t(\bar{\theta}^* - t)^2 - \sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2 \leqslant 0 \tag{9}$$

para cualquier estimación θ^*. No sería racional determinar el carácter bayesiano y minimax de las estimaciones, añadiendo simplemente a los primeros miembros el signo del paso límite ($\lim_{n\to\infty}$), ya que, por regla general, para las estimaciones a.n. de $\mathbf{M}_\theta(\theta^* - \theta)^2 \sim \sigma^2(\theta)/n$, los primeros miembros en (9) también convergerán hacia el cero. Por eso es natural examinar, digamos, la relación de los sumandos en (9). Teniendo en cuenta que más adelante se tratará principalmente de las estimaciones para las cuales $\mathbf{M}_\theta(\theta^* - \theta)^2$ tiene un orden de pequeñaz igual a $1/n$, se puede utilizar de un modo equivalente la definición siguiente.

Definición 3. La estimación θ_1^* se denomina *asintóticamente bayesiana o asintóticamente minimas,* si para cualquier otra estimación θ^* se cumple, respectivamente,

$$\lim_{n\to\infty} \sup\, [\mathbf{M}n(\theta_1^* - \theta) - \mathbf{M}n(\theta^* - \theta)^2] \leqslant 0,$$
$$\lim_{n\to\infty} \sup\, [\sup_{t\in\Gamma} \mathbf{M}_t n(\theta_1^* - t)^2 - \sup_{t\in\Gamma} \mathbf{M}_t n(\theta^* - t)^2] \leqslant 0.$$

Como veremos, la determinación de las estimaciones asintóticamente bayesianas y asintóticamente minimax es posible para suposiciones muy amplias.

En el *caso multidimensional* (cuando $\theta \in R^k$ es un vector) la propiedad 9) de la e.m.c., como hemos visto, se conserva, y la estimación

$$\theta_Q^* = \mathbf{M}(\theta/X)$$

minimizará

$$v(\theta^*) = \mathbf{M}(\theta^* - \theta)V(\theta^* - \theta)^T = \mathbf{M}\mathbf{M}_\theta(\theta^* - \theta)V(\theta^* - \theta)^T =$$
$$= \int \mathbf{M}_t(\theta^* - t)V(\theta^* - t)^T q(t)\lambda(dt)$$

para cualquier matriz V definida no negativamente o, que es lo mismo (véase el § 8), minimizará la desviación estándar $\theta^* - \theta$ promediada (con peso $q(t)$) en cualquier dirección $a \in R^k$.

Definición 4. La estimación θ_Q^* se llama *bayesiana* si para cualquier otra estimación θ^* y para cualquier matriz V definida no negativamente,

$$v(\theta_Q^*) \leqslant v(\theta^*).$$

La estimación θ_1^* se llama *asintóticamente bayesiana* si

$$\lim_{n\to\infty} \sup [nv(\theta_1^*) - nv(\theta_Q^*)] \leqslant 0.$$

Definición 5. La estimación $\bar{\theta}^*$ se denomina *minimax* si para cualquier otra estimación θ^* y para cualquier matriz V definida no negativamente,

$$\sup_{t\in\Gamma} \mathbf{M}_t(\bar{\theta}^* - t)V(\bar{\theta}^* - t)^T - \sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)V(\theta^* - t)^T \leqslant 0.$$

La estimación θ_1^* se denomina *asintóticamente minimax* si

$$\lim_{n\to\infty} \sup [\sup_{t\in\Gamma} \mathbf{M}_t n(\theta_1^* - t)V(\theta_1^* - t)^T - \sup_{t\in\Gamma} \mathbf{M}_t n(\bar{\theta}^* - t)V(\bar{\theta}^* - t)^T] \leqslant 0.$$

Concluyendo este párrafo señalaremos una vez más que las designaciones $\mathbf{M}_\theta S$, $\mathbf{P}_\theta(A)$, $f_\theta(x)$ en el caso bayesiano pueden ser consideradas, si es necesario, desde un nuevo punto de vista: como esperanzas matemáticas, probabilidades y densidades condicionales respecto a θ, o sea, como $\mathbf{M}(S/\theta)$, $\mathbf{P}(A/\theta)$ y $f(x/\theta)$, respectivamente.

§ 12. Estadísticas suficientes

En el párrafo anterior hemos examinado la cuestión acerca de la construcción de dos tipos de estimaciones óptimas: bayesianas y minimáx. En este párrafo introduciremos el concepto de estadística suficiente, que nos permitirá construir estimaciones eficientes, o sea, otro tipo de estimaciones óptimas destacadas en el § 8.

La noción de estadística suficiente desempeña un papel importante en la estadística matemática en general y en la teoría de las estimaciones en particular.

Convengamos en designar las estadísticas, o sea, las funciones medibles arbitrarias (escalares o vectoriales) de X, con el símbolo $S = S(X)$.

Sea $X \in \mathbf{P}_\theta$, $\mathbf{P}_\theta \in \mathscr{P} = \{\mathbf{P}_\theta\}$. Examinemos la distribución $\mathbf{P}_\theta(X \in B/S)$, $B \in \mathfrak{B}_{\mathscr{X}}^n$ que es condicional respecto a la variable aleatoria S y que ha sido engendrada por la distribución $\mathbf{P}_\theta$ en $\mathscr{X}^n$.

Definición 1. La estadística $S = S(X)$ se llama *suficiente para el parámetro* θ, si existe la variante de la distribución condicional $\mathbf{P}_\theta(X \in B/S)$ que no depende de θ.,

Sabemos que $\mathbf{P}_\theta(X \in B/S)$ es, para cada B, la e.m.c. y, por consiguiente, existe una función $\mathbf{P}(B/s)$ de Borel en s para cada B, tal que

$$\mathbf{P}_\theta(X \in B/S) = \mathbf{P}(B/S).$$

Podemos considerar (véase el § 10) que $\mathbf{P}(B/s)$, como función de B, es la *distribución condicional de las probabilidades, a condición de que* $S = s$. Esta distribución puede interpretarse como la *distribución de* X en *la superficie* $S(x) = s$.

Pero si S es una estádística suficiente , entonces dicha distribución ¡*no depende de* θ! Esto significa que el conocimiento del lugar donde se encuentra el punto muestral X en la superficie $S(x) = s$ no nos comunica ninguna información complementaria acerca del parámetro θ. (Pues está claro que nadie se dedicará a determinar el parámetro desconocido en el ejemplo 1 de la Introducción, con ayuda del lanzamiento de una moneda, puesto que la distribución del número de "caras" o "cruces" con tal lanzamiento no depende de θ en absoluto).

Esta circunstancia importante significa, a su vez, que toda la información acerca del parámetro θ está contenida en el valor de la estadística S. De aquí precisamente procede su nombre: estadística suficiente. Hablando en términos generales, el conocimiento de $S(X)$ es *suficiente* para construir el parámetro θ, pero los demás datos contenidos en la muestra X son inútiles.

Ejemplo 1. Sea $X \in \Pi_\lambda$. Demostremos que la estadística $S = n\bar{x} = \sum_{i=1}^{n} x_i$ es suficiente para el parámetro de la ley de Poisson λ. Debemos convencernos de que la distribución de la posición del punto X en la superficie $\sum_{i=1}^{n} x_i = s$ (s es un número entero) no depende de λ. En vista de que $\mathbf{P}(X = \mathbf{x}, \sum x_i = s) = \mathbf{P}(X = \mathbf{x})$ cuando $\sum_{i=1}^{n} x_i = s$, entonces

$$\mathbf{P}(X = x/n\bar{x} = s) = \begin{cases} \dfrac{\mathbf{P}(x_1 = x_1, \ldots, x_n = x_n}{\mathbf{P}(n\bar{x} = s)} & \text{si } \sum_{i=1}^{n} x_i = s, \\ 0 & \text{si } \sum_{i=1}^{n} x_i \neq s. \end{cases}$$

Como x_i son independientes, $\sum_{i=1}^{n} x_i \in \Pi_{n\lambda}$, el segundo miembro de (1) es igual a

$$\left(e^{-n\lambda} \frac{(n\lambda)^s}{s!}\right)^{-1} \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda x_i}{x_i!} = \frac{s!}{n^s \prod_{i=1}^{n} x_i!}.$$

Ahora bien, la distribución de X, que es condicional cuando $S = s$, coincide con la distribución polinomial $\mathbf{B}_p^s$ (véase el § 2) con n casos equiprobables (o sea, con el vector de probabilidades $p = (1/n, \ldots, 1/n)$) y con s pruebas independientes. Es evidente que la disribución no depende de λ, así que $S = n\bar{x}$ es una estadística suficiente para λ.

El concepto de estadística suficiente fue introducido en 1922 por Fisher. El siguiente teorema de Neyman — Fisher lleva el nombre de teorema de factorización y establece un criterio elemental de existencia de la estadística suficiente.

Supongamos que ha sido cumplida la condición (A_μ) de existencia de la densidad $f_\theta(x) = \dfrac{d\mathbf{P}_\theta}{d\mu}(x)$.

Teorema 1. *Para que S sea una estadística suficiente para θ, es necesario y suficiente que la función de verosimilitud* $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$ *sea representable en la forma*

$$f_\theta(x) = \psi(S(x), \theta)h(x) \quad c.s.[\mu^n], \tag{2}$$

donde cada una de las funciones $\psi \geqslant 0$ *y* $h \geqslant 0$ *depende sólo de sus propios argumentos,* $\psi(s, \theta)$ *es medible en s, y* $h(x)$, *en x.*

Por supuesto que la representación (2) no es unívoca. Sus componentes han sido determinados con una exactitud de hasta una función positiva arbitraria de $S(\overline{x})$.

En el ejemplo anteriormente examinado, con la distribución de Poisson,

$$f_\lambda(x) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{n\overline{x}}\prod_{i=1}^{n}\frac{1}{x_i!}, \quad n\overline{x} = \sum_{i=1}^{n} x_i,$$

así que podemos, para $S = n\overline{x}$, poner

$$\psi(S, \lambda) = e^{-n\lambda}\lambda^S h, \quad h(\overline{x}) = \prod_{i=1}^{n}\frac{1}{x_i!}.$$

De aquí, en virtud del teorema 1, resultará que $S = n\overline{x}$ es una estadística suficiente.

La demostración del teorema 1 aquí sólo se da para dos casos particulares más importantes: para el caso discreto y para el caso "suave". En el caso general, la demostración del teorema de Neyman — Fisher se da en el Suplemento IV.

En el *caso discreto,* μ es la medida de cálculo en el conjunto numerable $\mathscr{X}$ de los posibles valores de x_1 y, por lo tanto, $f_\theta(x) = \mathbf{P}_\theta(x_1 = x)$, $x \in \mathscr{X}$. Supongamos que al principio ha sido cumplida (2). Entonces, para el punto registrado $x \in \mathscr{X}^n$,

$$\mathbf{P}_\theta(X = x/S(X) = S(x)) = \frac{\mathbf{P}_\theta(X = x,\ S(X) = S(x))}{\mathbf{P}_\theta(S(X) = S(x))}. \tag{3}$$

Como $\{X = x,\ S(X) = S(x)\} = \{X = x\}$, el segundo miembro de (3) es igual a

$$\frac{\mathbf{P}_\theta(X = x)}{\mathbf{P}_\theta(S(X) = S(x))} = \frac{f_\theta(x)}{\sum\limits_{Y:S(y)=S(x)} f_\theta(y)} =$$

$$= \frac{\psi(S(x),\ \theta)\ h(x)}{\sum\limits_{y:S(y)=S(x)} \psi(S(x),\ \theta)h(y)} = \frac{h(x)}{\sum\limits_{y:S(y)=S(x)} h(y)}.$$

Ahora bien, $\mathbf{P}_\theta(X = x/S(X) = S(x)$ no depende de θ.

Al contrario, si el primer miembro de (3) no depende de θ, entonces, designándolo por $h(x)$, de (3) obtenemos $\mathbf{P}_\theta(X = x) = f_\theta(x) = \mathbf{P}_\theta(X = x;\ S(X) = S(x)) = h(x)\ \mathbf{P}_\theta(S(X) = S(x))$, donde $\mathbf{P}_\theta(S(X) = S(x)) = \psi(S(x),\ \theta)$ depende solamente de $S(x)$ y de θ. ◁

De un modo algo más complicado el teorema 1 también se demuestra en otro importante caso particular, o sea, en el caso "*suave*" cuando μ es la medida de Lebesgue en R, y la estadística $S(X)$ se supone que es función suave de X, es decir, una función tal que existe la sustitución de las variables $y_1 = S(x)$, $y_2 = y_2(x)$, ..., $y_n = y_n(x)$, resoluble respecto a $x_i = x_i(y_1, \ldots, y_n)$, con un jacobiano distinto del cero $J = \left|\frac{\partial x_i}{\partial y_j}\right| \neq 0$. En este caso, como es sabido de las fórmulas del análisis clásico sobre la sustitución de la variable en la integral, la densidad de la variable aleatoria $Y = (S(X), y_2(X)), \ldots, y_n(X))$ será igual a

$$g_\theta(y) = f_\theta(x)|J|,\ y = (y_1, \ldots, y_n).$$

La densidad de la variable aleatoria $y_1(X) = S(X)$ será igual a

$$g_\theta^{(1)}(y_1) = \int\limits_{R^{n-1}} g_\theta(y)dy_2 \ldots dy_n = \int\limits_{R^{n-1}} f_\theta(x)|J|dy_2 \ldots dy_n,$$

y la condicional de Y, a condición de que $S(X) = s$, será, por consiguiente, determinada por la relación

$$\varphi(y/s) = \frac{g_\theta(y)}{g_\theta^{(1)}(s)} = \frac{f_\theta(x)|J|}{g_\theta^{(1)}(s)} \text{ para } y_1 = s.$$

Después de estas observaciones preliminares, la demostración del teorema 1 para el caso "suave" se desarrolla al igual que para el caso discreto. En efecto, si se ha cumplido (2), entonces

$$\varphi(y/s) = \frac{\psi(s,\ \theta)h(x)|J|}{\int\limits_{R^{n-1}} \psi(s,\ \theta)h(x)|J|dy_2 \ldots dy_n}.$$

En esta relación, $\psi(s, \theta)$ se reduce. Esto significa que la distribución de Y, condicional a condición de que $S(X) = s$, y, por lo tanto, también la distribución de X no depende de θ,

Al contrario, si $\varphi(y/s)$ no depende de θ, entonces

$$f_\theta(x) = \frac{\varphi(y/s) g_\theta^{(1)}(s)}{|J|} \text{ cuando } s = S(x).$$

Esto significa que (2) se cumple cuando $\psi(s, \theta) = g_\theta^{(1)}(s)$, $h(x) = \varphi(y/s)/|J|$. ◁

Ejemplo 2. Sea $X \in \Phi_{\alpha,\sigma^2}$. Aquí el parámetro $\theta = (\alpha, \sigma^2)$ es bidimensional. Tenemos

$$f_\theta(X) = \prod_{i=1}^{n} \frac{-1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\alpha)^2}{2\sigma^2}} = \sigma^{-n}(2\pi)^{-n/2}\exp\left\{\frac{\sum(x_i-\alpha)^2}{2\sigma^2}\right\} =$$

$$= \sigma^{-n}\exp\left\{-\frac{\sum x_i^2 - 2\alpha n\bar{x} + n\alpha^2}{2\sigma^2}\right\}(2\pi)^{-n/2}.$$

Poniendo $S = (S_1, S_2)$, $S_1 = n\bar{x}$, $S_2 = \sum_{i=1}^{n} x_i^2$, obtenemos la representación (2), donde

$$\psi(S, \theta) = \sigma^{-n}\exp\left\{-\frac{S_2 - 2\alpha S_1 + \alpha^2}{2\sigma^2}\right\}, \quad h(X) = (2\pi)^{-n/2}.$$

Aquí podríamos, desde luego, atribuir el factor $(2\pi)^{-n/2}$ también a la función ψ, poniendo $h(X) = 1$.

Ahora bien, hemos obtenido que la estadística (S_1, S_2) es una estadística vectorial suficiente para (α, σ^2). De toda la información contenida en la muestra nos es suficiente saber $\bar{x}$ y $\sum x_i^2$.

Proponemos al lector hallar las estadísticas suficientes para todas las familias de distribuciones citadas en el § 2.

Concentraremos la atención tan sólo en una de estas familias.

Ejemplo 3. Sea $X \in U_{0,\theta}$. Aquí la condición (A_μ) se cumple con respecto a la medida de Lebesgue y

$$f_\theta(X) = \begin{cases} \theta^{-n} \text{ si } 0 \leqslant x_i \leqslant \theta \text{ cuando todos } i = 1, \ldots, n \\ 0 \text{ en el caso contrario.} \end{cases}$$

Sea $x_{(1)} = \text{mín } x_i$, $x_{(n)} = \text{máx } x_i$. Entonces, como hemos visto en el ejemplo 6.5, la función $f_\theta(X)$ puede ser escrita en forma de $f_\theta(X) = \psi(x_{(n)}, \theta)h(X)$, donde

$$h(X) = \begin{cases} 1 \text{ si } x_{(1)} \geqslant 0, \\ 0 \text{ en el caso contrario,} \end{cases}$$

$$\psi(s, \theta) = \begin{cases} \theta^{-n} \text{ para } s \leqslant \theta, \\ 0 \text{ en el caso contrario.} \end{cases}$$

Esto significa que $S(X) = x_{(n)}$ es una estadística suficiente para θ.

Análogamente el lector puede convencerse de que para la muestra $X \in \mathbf{U}_{\theta, 1+\theta}$, como estadística suficiente para el parámetro θ, sirve la estadística bidimensional $S(X) = (x_{(1)}, x_{(n)})$. Asimismo será la estadística suficiente para el parámetro bidimensional $\theta = (a, b)$ cuando la muestra ha sido extraída de la distribución $\mathbf{U}_{a,b}$.

Citaremos dos corolarios del teorema 1.

Corolario 1. *Si S es una estadística suficiente para θ, la estimación de verosimilitud máxima depende únicamente de S.*

Mejor dicho, la e.v.m. $\hat{\theta}^*$ no depende de X cuando se ha registrado $S(X)$.

Este corolario es evidente, ya que la e.v.m. es un valor de θ para el cual se alcanza el máximo de $f_\theta(X) = \psi(S(X), \theta)h(X)$ o bien, que es lo mismo, el máximo de $\psi(S(X), \theta)$.

Corolario 2. *Si S es una estadística suficiente y la función φ es tal que la aplicación $u = \varphi(v)$ es biunívoca y medible en ambas direcciones, entonces $S_1 = \varphi(S)$ también será una estadística suficiente.*

Este corolario también es evidente, puesto que $\psi(S, \theta)$ en (2) puede escribirse en forma de $\psi(\varphi^{-1}(S_1), \theta) = \psi_1(S_1, \theta)$.

También es válido un criterio más de suficiencia de la estadística S.

Teorema 2. *La estadística S es suficiente para θ si y sólo si para toda distribución a priori $\mathbf{Q}$ del parámetro θ la distribución a posteriori $\mathbf{Q}_X$ depende de X tan sólo a tráves de $S(X)$ (o sea, permanece invariable en la superficie de $S(X) = S$).*

Demostración. Supongamos que S es una estadística suficiente y que $q(t)$ es la densidad $\mathbf{Q}$ respecto a cualquier medida λ. Entonces, la densidad a posteriori $q(t/X)$ respecto a dicha medida, según la fórmula de Bayes será igual a

$$q(t/X) = \frac{f_t(X)q(t)}{\int f_u(X)q(u)\lambda(du)} = \frac{\psi(S(X), t)q(t)}{\int \psi(S(X), u)q(u)\lambda(du)}.$$

Demostremos ahora la afirmación inversa del teorema. Escojamos una distribución a priori de modo que $q(t) > 0$ en todas partes sobre Θ y para todos t

$$f_t(X) = \frac{q(t/X)f(X)}{q(t)}, \quad f(X) = \int f_u(X)q(u)\lambda(du).$$

Si $q(t/X) = g(t, S(X))$, entonces, poniendo $\psi(s, t) = g(t, s)/q(t)$, $h(X) = f(X)$, obtenemos la representación (2). ◁

Corolario 3. *Si S es una estadística suficiente, todas las estimaciones*

bayesianas y las estimaciones minimax definidas con ayuda del teorema 11.2 dependen únicamente de S.

En adelante obtendremos muchas otras confirmaciones de que la estadística suficiente S contiene la información completa acerca de θ.

§ 13. Estadísticas suficientes mínimas

Examinemos ahora la cuestión acerca de la elección de las características suficientes. Claro está que el número de éstas puede ser muy grande. Por ejemplo, la estadística $S(X) \equiv X$ siempre es evidentemente suficiente. La misma se llama estadística suficiente *trivial*. Sin embargo, estamos interesados (posteriormente será aclarado el porqué) en estadísticas más "económicas". Resulta que no siempre, ni micho menos, se pueden construir estadísticas suficientes que sean mucho más "económicas" que la estadística suficiente trivial. Volveremos a esta cuestión después que determinemos más exactamente los conceptos relacionados con la "economía" de las características suficiente... Para esto, introduzcamos en el conjunto de todas las características suficientes (para cierto parámetro θ), un orden parcial.

Definición 1. Diremos que la característica S_1 *está subordinada* a S_2 si S_1 es una función medible de S_2: $S_1 = \varphi(S_2)$.

Esta relación significa precisamente que S_1 es más "económica" que S_2.

Definición 2. Si S_1 está subordinada a S_2, y S_2 está subordinada a S_1, las estadísticas S_1 y S_2 se denominan *equivalentes*.

Evidentemente, S_1 es equivalente a S_2 si y sólo si $S_1 = \varphi(S_2)$ y φ es una aplicación biunívoca medible en ambas direcciones.

Definición 3. La estadística suficiente S_0 se denomina *mínima* si está subordinada a cualquier otra estadística suficiente S.

La estadística suficiente mínima es la más económica. Si hemos construido la estadística suficiente mínima S, entonces, siempre que se conserve la propiedad de suficiencia, será imposible la reducción ulterior de los datos en comparación con S. Los demás datos contenidos en la muestra pueden considerarse como engendrados por cierto mecanismo aleatorio no dependiente de θ, y ellos no proporcionan ninguna información acerca de θ.

Los conceptos introducidos, al igual que el concepto inicial de estadística suficiente, pueden exponerse, de forma ligeramente generalizada, en el lenguaje de las σ-álgebras, que en una serie de casos resulta más cómodo y evidente. Al principio —en la definición 1 del párrafo precedente— la distribución condicional $\mathbf{P}_\theta(X \in B/S)$ se puede sustituir por la distribución condicional $\mathbf{P}_\theta(X \in B/\mathfrak{U})$ respecto a la σ-subálgebra $\mathfrak{U} \subset \mathfrak{B}_1^n$ y la $\mathfrak{U}$ *σ-álgebra se puede llamar suficiente* si existe cierta variante $\mathbf{P}_\theta(X \in B/\mathfrak{U})$ que no depende de θ.

Con tal enfoque, el teorema de factorización se conserva si la función $\psi(S(X), \theta)$ es sustituida por la función $\psi(X, \theta)$ $\mathfrak{U}$-medible en X. La demostración de este teorema, expuesta en el Suplemento IV, prácticamente no se diferencia de la anterior.

La estadística suficiente ahora puede ser definida como una estadística S para la cual la σ-álgebra de $\sigma(S)$ será suficiente.

En el lenguaje de las σ—álgebras, la subordinación de las características suficientes (véase la definición 1) no exige que se introduzcan conceptos complementarios y coincide simplemente con el encaje de las σ-álgebras: S_1 está subordinada a S_2 si $\sigma(S_1) \subset \sigma(S_2)$. Ahora bien, S_1 es más económica que S_2 si la σ-álgebra de $\sigma(S_1)$ es más pobre que $\sigma(S_2)$. La equivalencia de S_1 y S_2 significa que $\sigma(S_1) = \sigma(S_2)$.

La *σ-álgebra suficiente mínima* de $\mathfrak{U}_0$ se define como una σ-álgebra que se encaja en cualquier σ-álgebra suficiente.

La *σ-álgebra suficiente mínima existe siempre.* Para convencerse de ello señalaremos previamente que, en virtud del teorema 2 del Suplemento IV, existe una distribución $\mathbf{Q}$ en Θ (además, discreta), tal que todas $\mathbf{P}_\theta$ son absolutamente continuas respecto a la distribución $\mathbf{P}_Q = \int \mathbf{P}_t \mathbf{Q}(dt)$.

Esto significa que $f_Q(X) = \int f_t(X)\mathbf{Q}(dt) > 0$ para todas X, o que de la igualdad $f_Q(X) = 0$ resulta $f_\theta(X) = 0$ para todos θ. En este caso se dice que $\mathbf{P}_Q$ *domina* la familia $\{\mathbf{P}_\theta\}$, así que podríamos adoptar $\mathbf{P}_Q$ como medida de μ. La densidad de la distribución $\mathbf{P}_\theta$ respecto a esta medida es igual a

$$\frac{d\mathbf{P}_\theta}{d\mathbf{P}_Q}(x) = \frac{f_\theta(x)}{f_Q(x)} \equiv r(x, \theta).$$

Está claro (compárese con el teorema 12.2) que si S es una estadística suficiente, $r(x, \theta)$ depende de x sólo a través de $S(x)$.

Teorema 1. *La σ-álgebra de $\mathfrak{U}_0 = \sigma(r(X, \theta); \theta \in \Theta)$ engendrada por las variables aleatorias $r(X, \theta) = f_\theta(X)/f_Q(X)$ para diferentes $\theta \in \Theta$, es una σ-álgebra suficiente mínima.*

La demostración del teorema es muy simple. La suficiencia de $\mathfrak{U}_0$ resulta del teorema de factorización y del hecho de que

$$f_\theta(X) = r(X, \theta) f_Q(X), \tag{1}$$

donde $f_Q(X)$ no depende de θ, y $r(X, \theta)$ es medible respecto a $\mathfrak{U}_0$.

Sea ahora $\mathfrak{U}$ cualquier σ-álgebra suficiente. Entonces $f_\theta(X) = \psi(X, \theta)h(X)$, donde la función $\psi(X, \theta)$ es $\mathfrak{U}$-medible. Examinemos la σ-álgebra de $\mathfrak{U}_\psi = \sigma(\psi(X, \theta), \theta \in \Theta) \subset \mathfrak{U}$. De la definición $r(X, \theta)$ se deduce que

$$r(X, \theta) = \frac{\psi(X, \theta)}{\int \psi(X, t)\mathbf{Q}(dt)}$$

y, por lo tanto, $\mathfrak{U}_0 \subset \mathfrak{U}_\psi \subset \mathfrak{U}$. ◁

Con este teorema y con el teorema 12.2 está estrechamente ligada otra afirmación útil. Examinemos el planteamiento bayesiano del problema cuando θ es una variable aleatoria con la distribución a priori $\mathbf{Q}$. Sea $q(t) > 0$ la densidad de esta distribución con respecto a la medida conveniente λ en Θ. Entonces la densidad a posteriori será igual a

$$q(t/X) = \frac{f_t(X)q(t)}{f_Q(X)} = r(X, t)q(t),$$

y, por consiguiente, la σ-álgebra suficiente mínima de $\mathfrak{U}_0$ puede considerarse como engendrada por la distribución a posteriori:

$$\mathfrak{U}_0 = \sigma(q(t/X);\ t \in \Theta).$$

Por regla general, la determinación de las distribuciones $\mathbf{Q}$ y $\mathbf{P}_Q$ que figuran en el teorema 1 no es difícil. Por ejemplo, si el portador N_{P_θ} de la distribución $\mathbf{P}_\theta$ no depende de θ, lo que tiene lugar para la mayoría de las distribuciones citadas en el § 2, se puede tomar $\mathbf{P}_Q = \mathbf{P}_{\theta_0}$ para cualquier $\theta_0 \in \Theta$.

Así pues, disponemos del teorema de existencia y del método eficaz para la construcción de las σ-álgebras suficientes mínimas *).

No obstante, las más de las veces para nosotros será más cómodo examinar las estadísticas. El fin principal de este párrafo consiste en determinar las estadísticas suficientes mínimas.

Ante todo, ¿de qué modo podemos comprobar que la estadística suficiente dada S_0 es mínima?

Una de las posibilidades consiste en la utilización del teorema 1. Si $\sigma(S_0)$ coincide con la σ-álgebra engendrada por $f_\theta(X)/f_Q(X)$, entonces S_0 es la estadística suficiente mínima.

Ejemplo 1. Hemos visto que la estadística $S = n\bar{x}$ es suficiente para el parámetro λ de la distribución de Poisson Π_λ. Ella será la estadística suficiente mínima, ya que $\sigma(S)$ coincide, evidentemente, con la σ-álgebra engendrada por $f_\lambda(X)/f_{\lambda_1}(X) = e^{n(\lambda_1 - \lambda)}(\lambda/\lambda_1)^S$ (aquí hemos tomado la distribución $\mathbf{Q}$ concentrada en el punto λ_1).

Ejemplo 2. Sea $X \in \mathbf{U}_{0,\theta}$. Entonces la estadística $S = x_{(n)} = \max x_i$ es la estadística suficiente mínima. En efecto, tomemos en calidad de $\mathbf{Q}$ cualquier distribución sobre $[0, \infty)$-con desindad $q(t) > 0$ para todos $t > 0$. Entonces

$$f_\theta(X) = \begin{cases} \theta^{-n}, & \theta \geqslant S, \\ 0, & \theta < S, \end{cases}$$

*) La existencia de la σ-álgebra suficiente mínima de $\mathfrak{U}_0$ también se puede establecer de otra manera, demostrando que $\mathfrak{U}_0$ es la intersección de todas las σ-álgebras suficientes completadas.

$$f_Q(X) = \int_0^\infty f_i(X)q(t)dt = \int_S^\infty t^{-n}q(t)dt > 0$$

para todas X. En este caso $S = \sup\{\theta: f_\theta(X)/f_Q(X) = 0\}$, lo cual significa que S es medible respecto a la σ-álgebra mínima de $\mathfrak{U}_0$, $\sigma(S) \subset \mathfrak{U}_0$ y que, por lo tanto, S es la estadística suficiente mínima.

Podemos indicar otro método de determinar las estadísticas suficientes mínimas, el cual también está relacionado con la función de verosimilitud. En efecto, toda estadística y, en particular, la estadística suficiente S engendra la partición del espacio muestral en clases de equivalencia, o sea, en conjuntos de los puntos x con iguales valores de $S(x)$.

Si S_1 está subordinada a S_2, o sea, $S_1 = \varphi(S_2)$, es evidente que para S_1 la partición será más grande, ya que las clases de equivalencia para S_2 se contienen en las de equivalencia para S_1. Ahora bien, a la estadística suficiente mínima le corresponde la "mayor" partición entre las particiones engendradas por las estadísticas suficientes.

Se pueden examinar simplemente las particiones del espacio en clases de equivalencia sin relacionarlas directamente con las estadísticas. Designemos por $D(x)$ la clase de equivalencia que contiene el punto x. Cada clase se define unívocamente por un punto cualquiera. *Llamaremos suficiente la partición en clases D si*

$$f_\theta(x) = \varphi(x, \theta)h(x), \tag{2}$$

donde $\varphi(x, \theta) = \varphi(x_\theta, \theta)$ *es constante para* $x \in D(x_0)$ (o sea, $\varphi(x, \theta) = \text{const}$ dentro de la clase de equivalencia). Si las clases $D(x)$ son definidas por las relaciones $S(x) = s$, del teorema 11.1 se desprende directamente que la estadística $S(x)$ es suficiente si y sólo si la partición en clases D es suficiente.

Examinemos ahora la partición construida del modo siguiente: tomemos el punto x_0 y declaremos que x pertenece a la clase $D(x_0)$ si la relación

$$\frac{f_\theta(x)}{f_\theta(x_0)} = h(x, x_0) \tag{3}$$

no depende de θ. Es evidente que con tal construcción, $D(x_1) = D(x_2) = D(x_0)$ si $x_1 \in D(x_0)$, $x_2 \in D(x_0)$, así que la regla (3) engendra la partición de todo el espacio en clases disjuntas.

Esta partición corresponde a la engendrada por la estadística suficiente mínima S.

En efecto, sea S la estadística suficiente mínima. Tomemos un punto arbitrario x_0. Entonces sobre la superficie $S(x) = S(x_0)$, la relación $f_\theta(x)/f_\theta(x_0)$ es igual a $h(x)/h(x_0)$ y, por consiguiente, no depende de Θ. Así pues, la partición en clases D es no menos grande que la partición para S.

Por otro lado, esta partición es suficiente. Efectivamente, podemos hacer que a cada superficie D le corresponda un punto cualquiera x_D de ella, a partir del cual la misma será definida unívocamente. Examinemos la función $x_0(x)$ que se define según la relación $x_0(x) = x_D$ si $x \in D$. Entonces, en virtud de (3), cuando $x \in D$,

$$f_\theta(x) = f_\theta(x_D)h(x, x_D) = f_\theta(x_0(x))h(x_1, x_0(x)), \tag{4}$$

que significa el cumplimiento de (2).

Los planteamientos efectuados no han sido del todo estrictos, ya que no los hemos relacionado con la cuestión acerca de la mensurabilidad de las funciones que forman parte de (4).

Lo dicho se puede resumir del modo siguiente. *Supongamos que se da una estadística* $S(X)$ *tal que* $S(x) = S(x_0)$ *si y sólo si la relación* (3) *no depende de* θ. *En este caso S es la estadística suficiente mínima.*

A distinción de los enfoques relacionados con el teorema 1, donde fueron examinadas las relaciones $f_\theta(x)/f_Q(x)$ o bien $f_\theta(x)/f_{\theta_1}(x)$ para diferentes θ y θ_1 (denominadas con frecuen-

cia relaciones de verosimilitud), la regla enunciada más arriba utiliza la relación $f_\theta(x)/f_\theta(x_0)$ para iguales valores del parámetro θ. En el ejemplo 1, por ejemplo, la relación

$$f_\lambda(x)/f_\lambda(x_0) = \Pi\lambda^{x_i - x_{i0}}x_{i0}!/x_i! = \lambda^{n(\bar{x} - \bar{x}_0)}\Pi x_{i0}!/x_i!$$

no dependerá de λ si y sólo si $\bar{x} = \bar{x}_0 = \frac{1}{n}\sum_{i=1}^{n} x_{i0}$, donde x_{i0} son las coordenadas del vector x_0. Esto es suficiente para sacar la conclusión de que $S(x) = \bar{x}$ es la estadística suficiente mínima.

Valiéndonos de la regla propuesta, examinemos ahora un ejemplo cuando no existen estadísticas suficientes " económicas". Antes que nada señalaremos que la serie variacional $S_V = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, construida según la muestra X, es siempre, evidentemente, la estadística suficiente, ya que $f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{k=1}^{n} f_\theta(x_{(k)})$. Esta estadística es "un poco más económica" que la propia muestra x. De aquí, en particular, se deduce que cualquier estadística suficiente mínima es invariante con respecto a la permutación de las coordenadas x_i en la muestra X.

Si la densidad $f_\theta(x)$ es simétrica, o sea, $f_\theta(-x) = f_\theta(x)$ para todos θ, es evidente que existirá una estadística suficiente, un poco más "económica", que representa la población $(x_1^2, \ldots, x_n^2)$ ordenada en función de su crecimiento y que designaremos por S_{V^2}.

Ejemplo 3. Si $X \in \mathbf{K}_{0,\sigma}$, o sea, si x_i tiene densidad de distribución de Cauchy con parámetro $\theta = \sigma$,

$$k_{0,\sigma}(x) = \frac{\sigma}{\pi(x^2 + \sigma^2)},$$

la estadística S_{V^2} será la estadística suficiente mínima.

En efecto, en este caso

$$f_\sigma(x) = \left(\frac{\sigma}{\pi}\right)^n \prod_{i=1}^{n} (x_i^2 + \sigma^2)^{-1},$$

así que

$$\frac{f_\sigma(x)}{f_\sigma(x_0)} = \prod_{i=1}^{n} \frac{x_{i0}^2 + \sigma^2}{x_i^2 + \sigma^2} \tag{5}$$

es la relación de dos polinomios de σ^2, la cual no depende de σ si y sólo si los coeficientes de las potencias correspondientes de σ^2 coinciden en el numerador y el denominador. Esto, a su vez, tiene lugar si y sólo si los conjuntos de "ceros" $[-x_{i0}^2]$ y $[-x_i^2]$ coinciden. Con otras palabras, para que (5) sea independiente de σ es necesario y suficiente que el punto $x^2 = (x_1^2, \ldots, x_n^2)$ tenga coordenadas que se distingan de las de x_0^2 tan sólo por la permutación de sus lugares. Esto precisamente significa que S_{V^2} es una estadística suficiente mínima.

De manera completamente análoga se puede demostrar que S_V es una estadística suficiente mínima para el parámetro α y, por lo tanto, para el parámetro $\theta = (\alpha, \sigma)$ de la distribución $\mathbf{K}_{\alpha,\sigma}$.

Otro ejemplo, en el que S_V será una estadística sifuciente mínima, se obtiene si se examina la familia

$$\mathbf{P}_{\alpha,\theta_1,\theta_2} = \alpha\mathbf{P}_{\theta_1} + (1 - \alpha)\mathbf{P}_{\theta_2}, \qquad \alpha \in [0, 1],$$

donde $\{\mathbf{P}_\theta\}$ es una familia exponencial (véase § 15, en calidad de $\mathbf{P}_\theta$ se puede tomar la distri-

bución normal o la distribución de Poisson) y donde al menos uno de los parámetros α, θ_1, θ_2 se desconoce.

Ahora demostremos un teorema que indica un método "estructural" simple de determinación de las estadísticas suficientes mínimas.

Para simplificar la exposición examinemos el caso del parámetro unidimensional θ.

Teorema 2. *Supongamos que la función de verosimilitud $f_\theta(x)$, para todas x como función de θ, es continua a la derecha (o a la izquierda). Entonces, si la estimación de v.m. $\hat{\theta}^*$ es única y la misma es una estadística suficiente, entonces $\hat{\theta}^*$ será la estadística suficiente mínima.*

Demostración. Sea S una estadística suficiente arbitraria. Demostraremos el teorema si mostramos que $\hat{\theta}^*$ es medible respecto a $\sigma(S)$ y, por lo tanto, $\hat{\theta}^*$ está subordinada a S.

En virtud del teorema de factorización,

$$f_\theta(x) = \psi(S(x), \theta)h(x) \text{ c.s.}[\mu^n], \tag{6}$$

donde $h(x)$ es la función medible en x, y $\psi(s, t)$ es continua (a la derecha o a la izquierda) en t y medible en s. Como $\mathbf{P}_\theta$ no variará si la densidad $f_\theta(x)$ cambia en el conjunto de la μ^n-medida 0, podemos considerar que (6) es válida para todos x.

En virtud de (6), el punto del máximo absoluto de $f_\theta(x)$ también es el punto del máximo absoluto para $\psi(S(x), \theta)$. Por eso, en virtud de la unicidad de $\hat{\theta}^*$,

$$\{\hat{\theta}^* < t\} = \{\sup_{\theta < t} \psi(S(X), \theta) > \sup_{\theta \geqslant t} \psi(S(X), \theta).$$

En vista de que $\psi(S(X), \theta)$, para cada $S(X)$, es continua en θ a la derecha (o a la izquierda), existe un conjunto numerable, denso en todas las partes, $\Theta_c = \{\theta_j\}_{j=1}^{\infty} \subset \Theta$ (igual para todos los $S(X)$) tal que

$$\sup_{\theta < t} \psi(S(X), \theta) = \sup_{\substack{\theta_j < t \\ \theta_j \in \Theta_c}} \psi(S(X), \theta_j). \tag{7}$$

Esa misma relación será válida para la región de $\theta \geqslant t$. Como $\psi(S(X), \theta_j)$ son medibles respecto a $\sigma(S)$, en virtud de (7), los valores de $\sup_{\theta < t} \psi(S, \theta)$ y $\sup_{\theta \geqslant t} \psi(S, \theta)$ serán variables aleatorias también medibles con respecto a $\sigma(S)$. Por consiguiente, $\{\theta^* < t\} \in \sigma(S)$, y el teorema ya está demostrado. ◁

En la condición de la afirmación citada, la condición de suficiencia de la e.v.m. $\hat{\theta}^*$ es esencial, puesto que la estimación de verosimilitud máxima $\hat{\theta}^*$, como tal, no es obligatoriamente una estimación suficiente. Es fácil

obtener un ejemplo respectivo examinando cualquier familia de distribuciones $\{\mathbf{P}_\theta\}$, con parámetro escalar θ y con estadística suficiente mínima vectorial S(cuya dimensión es mayor que 1). En este caso la estimación de verosimilitud máxima $\hat{\theta}^*$ también será escalar, así que la σ-álgebra de $\sigma(S)$ será más rica que $\sigma(\hat{\theta}^*)$ y, por lo tanto, la inclusión de $\sigma(S) \subset \sigma(\hat{\theta}^*)$, que se desprende de la minimalidad de S y de la suficiencia de $\hat{\theta}^*$, es imposible.

Ejemplo 4. Sea $X \in U_{\theta,1+\theta}$, $\Theta = R$. Entonces, como hemos visto en el ejemplo 6.4,

$$f_\theta(X) = \begin{cases} 1 \text{ para } \theta \leqslant x_{(1)} \leqslant x_{(n)} \leqslant 1 + \theta, \\ 0 \text{ en el caso contrario,} \end{cases}$$

así que $f_\theta(X)$ depende de X solamente a través de $x_{(1)}$ y $x_{(n)}$. Esto significa que $S = (x_{(1)}, x_{(n)})$ es una estadística suficiente. Ni una de las magnitudes $x_{(1)}$, $x_{(n)}$ por separado es una estadística suficiente. Eso lo demuestran las relaciones siguientes:

$$\mathbf{P}(x_{(1)} \geqslant u,\ x_{(n)} < v) = \prod_{i=1}^{n} \mathbf{P}(x_i \in [u,\ v)) =$$

$$= (v - u)^n \text{ cuando } u \geqslant \theta,\ v \leqslant 1 + \theta,\ v > u.$$

Por consiguiente, la densidad compatible de distribución de $(x_{(1)}, x_{(n)})$ será igual a

$$g(u,\ v) = \begin{cases} n(n-1)(v-u)^{n-2} & \text{cuando } u \geqslant \theta,\ v \leqslant 1 + \theta,\ v > u, \\ 0 & \text{en los demás casos.} \end{cases}$$

Seguidamente, $\mathbf{P}(x_{(1)} \geqslant u) = (1 + \theta - u)^n$ cuando $\theta \leqslant u \leqslant 1 + \theta$, así que la densidad de $x_{(1)}$ es igual a

$$g(u) = n(1 + \theta - u)^{n-1} \text{ cuando } \theta \leqslant u \leqslant 1 + \theta.$$

De aquí ya es fácil obtener que la densidad condicional $g(v/u)$ de la magnitud $x_{(n)}$, a condición de $x_{(1)} = u$ (y, por lo tanto, también la distribución condicional correspondiente), dependerá de θ. Esto significa que $x_{(1)}$ (al igual que $x_{(n)}$) por separado no son estadísticas suficientes. Como en calidad de e.v.m. $\hat{\theta}^*$ podemos tomar $\hat{\theta}^* = x_{(1)}$ (véase el ejemplo 6.4, por lo tanto, hemos demostrado que para la familia $U_{\theta,1+\theta}$, la e.v.m. $\hat{\theta}^*$ no es una estadística suficiente.

Mediante el teorema 1, el lector puede convencerse personalmente de que $S = (x_{(1)}, x_{(n)})$ es una estadística suficiente mínima para $U_{\theta,1+\theta}$.

La condición de suficiencia de $\hat{\theta}^*$ en el teorema 2 será cumplida automáticamente si suponemos que existe una estadística suficiente escalar (para un θ unidimensional) S_0, para la cual la función φ en la igualdad $\hat{\theta}^* = \varphi(S_0)$ será biunívoca (o sea, $\hat{\theta}^*$ y S_0 serán equivalentes).

§ 14. Construcción de estimaciones eficientes por medio de estadísticas suficientes. Estadísticas completas

Definición 1. La estimación θ^* se denomina *suficiente* si es una estadística suficiente.

1. Caso unidimensional. Supondremos aquí que θ es un parámetro escalar. Sea K_b la clase de todas las estimaciones θ^* con desplazamiento $b(\theta)$, así que $\theta^* \in K_b$ si $a(\theta) = \mathbf{M}_\theta\theta^* = \theta + b(\theta)$. Para $\theta^* \in K_b$ tenemos

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \mathbf{M}_\theta(\theta^* - a(\theta))^2 + (a(\theta) - \theta)^2 = \mathbf{D}_\theta\theta^* + b^2(\theta).$$

En este párrafo omitiremos, a veces, el índice θ de los símbolos $\mathbf{M}_\theta$, $\mathbf{D}_\theta$.

La siguiente afirmación fue obtenida independientemente por Blackwell, Rao y Kolmogórov.

Teorema 1. *Sea S una estadística suficiente, $\theta^* \in k_b$. Entonces la función $\theta_S^* = \mathbf{M}_\theta(\theta^*/S)$ es una estimación que posee las siguientes propiedades:*

1) $\theta_S^* \in K_b$,

2) *θ_S^* depende de la muestra tan sólo a través de $S(X)$,*

3) $\mathbf{M}_\theta(\theta_S^* - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta)^2$ *para todos θ.*

La última desigualdad se transforma en igualdad tan sólo si $\theta^* = \theta_S^*$ c.d. respecto a $\mathbf{P}_\theta$.

Con otras palabras, en la clase K_b, la aplicación de la operación $\mathbf{M}_\theta(\cdot/S)$ a θ^* mejora uniformemente la estimación θ^*.

Demostración. El hecho de que θ_S^* es una estimación, significa que θ_S^* no depende de θ y que es una función medible de X. Su independencia respecto a θ se desprende de las propiedades de las características estadísticas, ya que la distribución de X para una S registrada no depende de θ ($\mathbf{M}_\theta(\theta^*/S)$, para la estadística arbitraria S, hablando en general, depende de θ). Al mismo tiempo, en virtud de las propiedades de la e.m.c., θ_S^* es una función medible de S y, por lo tanto, también de X. Por consiguiente, θ_S^* es la estimación que satisface la propiedad 2) del teorema.

La igualdad

$$\mathbf{M}_\theta\theta_S^* = \mathbf{M}_\theta\mathbf{M}_\theta(\theta^*/S) = \mathbf{M}_\theta\theta^*,$$

que demuestra que $\theta_S^* \in K_b$, también se deduce directamente de las propiedades de la e.m.c. Seguidamente,

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \mathbf{M}_\theta(\theta^* - \theta \pm \theta_S^*)^2 = \mathbf{M}_\theta(\theta_S^* - \theta)^2 + \mathbf{M}_\theta(\theta^* - \theta_S^*)^2 + \\ + 2\mathbf{M}_\theta(\theta_S^* - \theta)(\theta^* - \theta_S^*).$$

Utilizando de nuevo las propiedades de la e.m.c., obtenemos

$$\mathbf{M}_\theta(\theta_S^* - \theta)(\theta^* - \theta_S^*) = \mathbf{M}_\theta\mathbf{M}_\theta[(\theta_S^* - \theta)(\theta^* - \theta_S^*)/S] =$$

$$= \mathbf{M}_\theta[(\theta_S^* - \theta)\mathbf{M}_\theta(\theta^* - \theta_S^*/S)] = 0$$

y, por consiguiente,

$$\mathbf{M}_\theta(\theta^* - \theta)^2 - \mathbf{M}_\theta(\theta_S^* - \theta)^2 + \mathbf{M}_\theta(\theta^* - \theta_S^*)^2. \quad \triangleleft$$

En realidad, la desigualdad 3) del teorema 1 se puede obtener directamente de la propiedad de la e.m.c., $(\mathbf{M}(\xi/S))^2 \leqslant \mathbf{M}(\xi^2/S)$, ya que entonces

$$(\theta_S^* - \theta)^2 = [\mathbf{M}_\theta(\theta^* - \theta)/S)]^2 \leqslant \mathbf{M}_\theta[(\theta^* - \theta)^2/S],$$

$$\mathbf{M}_\theta(\theta_S^* - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta)^2.$$

El hecho expuesto en el teorema 1 puede interpretarse del modo siguiente. Supongamos que S y T son dos estadísticas suficientes, $\theta^* = \varphi(T)$ y S está subordinada a T, entonces $\mathbf{M}_\theta(\theta_S^* - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta)^2$.

Con otras palabras, cuanto más "económica" sea la estadística suficiente S (o cuanto más pobre sea la σ-álgebra correspondiente), tanto mejores serán las estimaciones θ_S. Así pues, para construir las estimaciones óptimas debemos buscar las estadísticas suficientes *mínimas* (o las σ-álgebras mínimas). En este caso, en calidad de estimaciones iniciales θ^* también pueden figurar estimaciones "malas" que no poseen, por ejemplo, incluso propiedad de validez. En este sentido es aleccionador el siguiente

Ejemplo 1. Sea $X \in \Pi_\lambda$. La estimación $\lambda^* = x_1$, evidentemente, no está desplazada $\mathbf{M}\lambda^* = \mathbf{M}x_1 = \lambda$ ($b\lambda = 0$) y no es válida, ya que no depende de n. La estadística suficiente mínima de λ es la estadística $S = n\overline{x} = \sum x_i$. Del ejemplo 12.1 se deduce que la distribución x_1 condicional respecto a S es la distribución $\mathbf{B}_{1/n}^S$ en el esquema de Bernoulli, con una probabilidad de éxito igual a $1/n$:

$$\mathbf{P}(x_1 = k/S = s) = C_s^k\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{s-k}.$$

Por consiguiente,

$$\lambda_S^* = \mathbf{M}(x_1/S) = \sum_{k=1}^{S} kC_S^k\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{S-k} = \frac{S}{n} = \overline{x}.$$

En uno de los ejemplos ulteriores demostraremos que $\overline{x}$ es una estimación eficiente.

2. Caso multidimensional. Ahora obtendremos los análogos del teorema 1 para el caso multidimensional cuando θ y θ^* son vectores de R^k.

Al igual que en el caso unidimensional, el vector $b(\theta) = \mathbf{M}_\theta\theta^* - \theta$ será

el desplazamiento de la estimación θ^*, y por K_b designaremos la clase de todas las estimaciones con desplazamiento b.

Teorema 1A. *Sea S una estadística suficiente y $\theta^* \in K_b$. Entonces la estimación $\theta_S^* = \mathbf{M}_\theta(\theta^*/S)$ posee las propiedades*

1) $\theta_S^* \in K_b$.

2) *θ_S^* depende exclusivamente de $S(X)$,*

3) *la dispersión estándar de θ_S^* no supera la dispersión estándar de θ^* o bien, que es lo mismo, para cualquier vector $a \in R^k$*

$$\mathbf{M}_\theta(\theta_S^* - \theta,\ a)^2 \leqslant \mathbf{M}_\theta(\theta^* - \theta,\ a)^2. \tag{1}$$

Aquí, la igualdad (para todos los valores de a) es posible únicamente en el caso de $\theta^* = \theta_S^*$ c.d. respecto a $\mathbf{P}_\theta$.

Demostración. Las primeras dos afirmaciones son evidentes. Las desigualdades (1) se deducen del teorema 1, puesto que todo se reduce al examen de las estimaciones unidimensionales (θ^*, a) del parámetro (θ, a), y $\mathbf{M}_\theta[(\theta^*, a)/S] = (\theta_S^*, a)$. Si en (1), para todos los valores de a es válida esa igualdad, entonces, para cada a tendremos $(\theta_S^*, a) = (\theta^*, a)$ c.d. Esto precisamente significa que $\theta_S^* = \theta^*$ c.d. ◁

Ahora bien, en el caso multidimensional, las estadísticas suficientes desempeñan el mismo papel: la forma cuadrática $\sum \sigma_{ij} a_i a_j$, donde $\sigma^2 = \|\sigma_{ij}\|$ es la matriz de segundos momentos para $\theta_S^* - \theta$, será tanto menor cuanto menor sea la σ-álgebra de $\sigma(S)$ engendrada por S.

3. Estadísticas completas y estimaciones eficientes. Ahora citaremos un criterio muy simple del inmejoramiento de las estimaciones, basado en el concepto de plenitud de la característica S. Designemos por l la dimensión de la característica S. Esta suele ser mayor que la dimensión k del parámetro θ o igual a ésta.

Para dos funciones medibles $f_1(s)$ y $f_2(s)$:$R^l \to R^k$ escribiremos $f_1(s) = f_2(s)$ c.d. $[\mathscr{P}]$, donde $\mathscr{P}$ es la familia de distribuciones en $(R^l, \mathfrak{B}^l)$ si $f_1(s) = f_2(s)$ en todas las partes excepto el conjunto N tal que $\mathbf{P}(N) = 0$ para todas $\mathbf{P} \in \mathscr{P}$

Definición 2. La familia de distribuciones $\mathscr{G} = \{\mathbf{G}_\theta\}$ en $(R^l, \mathfrak{B}^l)$, que dependen del parámetro k-dimensional $\theta \in \Theta \subset R^k$, se llama *completa* si la igualdad

$$\int y(s)\, \mathbf{G}_\theta(ds) = 0 \text{ cuando todos } \theta \in \Theta \tag{2}$$

conduce a $y(s) = 0$ c.d. $[\mathscr{G}]$. La ecuación (2) se examina en la clase de funciones y: $R^l \to R^k$ para las cuales existe la integral (2).

Definición 3. La estadística S se denomina *completa* si la familia $\mathscr{G}$ de sus distribuciones $\mathbf{G}_\theta$, inducidas por la distribución $\mathbf{P}_\theta$ en $(\mathscr{X}^n, \mathfrak{B}_{\mathscr{X}}^n)$, es completa.

La ecuación (2) para las estadísticas puede ser escrita en forma de $\mathbf{M}_\theta y(S) = 0$ para todos $\theta \in \Theta \subset R^k$.

Teorema 2. *La estadística S es completa si y sólo si para cualquier $b_0(\theta)$, la $\sigma(S)$-medible* [*]) *estimación θ^* es única en la clase de todas las $\sigma(S)$-medibles estimaciones de K_{b_0}.*

Si la $\sigma(S)$-medible estimación es única en K_{b_0}, entonces las $\sigma(S)$-medibles estimaciones también poseerán la propiedad de unicidad en cualquier otra clase K_b.

La *demostración* de esta afirmación es casi evidente, ya que la existencia de dos $\sigma(S)$-medibles estimaciones $\theta_1^* = \varphi_1(S)$ y $\theta_2^* = \varphi_2(S)$ en K_{b_0} significa que $\int \varphi_i(s)\mathbf{G}_\theta(ds) = b_0(\theta)$, $i = 1, 2$,

$$\int [\varphi_1(s) - \varphi_2(s)]\mathbf{G}_\theta(ds) = 0 \text{ para todos } \theta \in \Theta,$$

así que la plenitud de S conduce a $\varphi_1(s) = \varphi_2(s)$ c.d. [$\mathscr{S}$]. Al contrario, sea $\int y(s)\mathbf{G}_\theta(ds) = 0$ para todos $\theta \in \Theta$, $\theta_1^* = \varphi_1(s) \in K_b$. Entonces $\theta_2^* = \varphi_1(s) + y(s) \in K_b$, y la unicidad de la $\sigma(S)$-medible estimación significa que $y(s) = 0$ c.d. [$\mathscr{S}$]. ◁

Teorema 3. *Si la estadística suficiente S es completa, y $\theta^* \in K_b$, entonces la estimación $\theta_S^* = \mathbf{M}_\theta(\theta^*/S)$ es la estimación eficiente única en K_b.*

Este teorema nos ofrece criterios suficientemente simples de eficacia de las estimaciones.

Demostración. En virtud del teorema (2), la $\sigma(S)$-medible estimación en la clase K_b es única.

Sea θ^{**} cualquier otra estimación de K_b. Entonces $\theta_S^{**} = \mathbf{M}_\theta(\theta^{**}/S) \in K_b$ y, por lo tanto, $\Theta_S^* = \theta_S^*$ c.d. [$\mathscr{S}$]. De aquí y del teorema 1 se desprende que

$$\mathbf{M}_\theta(\theta_S^* - \theta)^2 = \mathbf{M}_\theta(\theta_S^{**} - \theta)^2 \leqslant \mathbf{M}_\theta(\theta^{**} - \theta)^2,$$

y la igualdad es posible únicamente para $\theta^{**} = \theta_S^*$ c.s. ◁

Corolario 1. *Si S es una estadística suficiente completa, y θ^* es una estimación no desplazada, entonces θ_S^* es una estimación eficiente y es la única.*

Ejemplo 2. En el ejemplo 1, con distribución de Poisson, hemos obtenido que para $\lambda^* = x_1$

$$\lambda_S^* = \mathbf{M}_\lambda(x_1/S) = \bar{x},$$

donde $S = n\bar{x}$. Mostremos que S es una estadística completa y, por consiguiente, $\bar{x}$ es una estimación suficiente. La ecuación (2) para la estadística

---

[*]) O sea, medible respecto a la σ-álgebra de $\sigma(S)$ engendrada por S y, por lo tanto, representable en forma de $\varphi(S)$, donde φ es la función de Borel.

S tiene la forma

$$\sum_{k=0}^{\infty} v(k) e^{-n\lambda} \frac{(n\lambda)^k}{k!} = 0 \text{ cuando todos } \lambda \geqslant 0,$$

o, que es lo mismo,

$$v(z) = \sum v(k) \frac{z^k}{k!} = 0 \text{ para todos } z \geqslant 0. \tag{3}$$

Es evidente que esto conduce a $v(k) = 0$, ya que de la convergencia de la serie (3), digamos, cuando $z = 1$ se deduce que $v(z)$ es analítica cuando $|z| < 1$ y es idénticamente igual a 0. Por consiguiente, los coeficientes $v(k)$ de su desarrollo en serie son iguales a 0.

Ejemplo 3. Sea $X \in \mathbf{U}_{0,\theta}$. Mostremos que la estadística $S = x_{(n)} = \max_{i \leqslant n} x_i$ es completa. La suficiencia (y minimización) de S ha sido establecida en el ejemplo 13.2. La distribución de S se define por la igualdad

$$\mathbf{P}(S < s) = (s/\theta)^n, \quad 0 \leqslant s \leqslant \theta,$$

así que S tiene una densidad igual a $ns^{n-1}\theta^{-n}$ cuando $s \in [0, \theta]$. En este caso la ecuación (2) tiene la forma

$$\int_0^{\theta} v(s) \frac{ns^{n-1}}{\theta^n} ds = 0 \text{ cuando } \theta \in (0, \infty).$$

De la igualdad $\int_0^{\theta} v(s) s^{n-1} ds = 0$ para todos θ resulta, evidentemente, que $v(s)s^{n-1} = 0$, $v(s) = 0$ c.d.

Le proponemos al lector que verifique si son completas las estadísticas suficientes para otras familias paramétricas y, en particular, que determine si $\alpha^* = \frac{1}{\bar{x}} \left(1 - \frac{1}{n}\right)$ es la estimación eficiente única del parámetro α de la familia $\Gamma_{\alpha,1}$ (véase § 2).

Señalemos ahora que el teorema 3 muestra la existencia de relaciones entre los conceptos de amplitud y minimización. En este aspecto es válida la afirmación siguiente, que da, junto con los teoremas del § 13, el criterio de minimización de las estadísticas suficientes.

Teorema 4. *Cualquier característica suficiente completa S es una estadística suficiente mínima.*

Demostración. Sea $\mathfrak{U}_0$ una σ-álgebra suficiente mínima (según el teorema 13.1, ésta existe). Supongamos que $\mathbf{M}_\theta S$ existe y examinemos la función

$\psi = S - \mathbf{M}_\theta(S/\mathfrak{U}_0)$. Como $\mathfrak{U}_0 \subset \sigma(S)$, entonces ψ será $\sigma(S)$-medible, así que $\psi = \psi(S)$. Designemos por $\mathbf{G}_\theta$ la distribución de S. Entonces es evidente que para todos θ, $\mathbf{M}_\theta\psi(S) = 0$ o, que es lo mismo,

$$\int \psi(s)\mathbf{G}_\theta(ds) = 0 \text{ para todos } \theta \in \Theta.$$

De aquí, en virtud de la amplitud de S resulta que $\psi(s) = 0$ c.s. $[\mathscr{S}]$, $\mathscr{S} = \{\mathbf{G}_\theta\}$. Esto significa que $S = \mathbf{M}_\theta(S/\mathfrak{U}_0)$ c.d. $[\mathscr{S}]$ y, por lo tanto, S es medible respecto a*) $\mathfrak{U}_0$, $\sigma(S) = \mathfrak{U}_0$.

Si $\mathbf{M}_\theta S$ no existe, es necesario, en vez de S, examinar la estadística arctg S, la cual es, evidentemente, equivalente a S en cuanto a las propiedades de suficiencia, amplitud y minimización. ◁

Señalemos que la afirmación inversa no es cierta: la *estadística suficiente mínima no es obligatoriamente completa.* Los ejemplos respectivos se obtienen fácilmente en los casos en que la dimensión l de la estadística es mayor que la dimensión k del parámetro θ. Por ejemplo, en el § 13 hemos visto que la densidad compatible de la estadística suficiente mínima $S = (x_{(1)}, x_{(n)})$ para la familia $\mathbf{U}_{\theta, 1+\theta}$ es igual a

$$g_\theta(u, v) = \begin{cases} n(n-1)(v-u)^{n-2} & \text{cuando } u \geqslant \theta,\ v \leqslant 1+\theta,\ v > u, \\ 0 & \text{en los demás casos.} \end{cases}$$

Si se toma la función $y(u, v) = \varphi(v - u)$ y se hace la transformación ortogonal $(v-u)/\sqrt{2} = t$, $(v+u)/\sqrt{2} = z$, la integral en (2) por el triángulo $u \geqslant \theta$, $v \leqslant 1+\theta$, $v > u$) será igual a

$$\int y(u, v)g_\theta(u, v)du\,dv = n(n-1)\int_0^1 \varphi(x)x^{n-2}(1-x)dx.$$

Es evidente que la integral en el segundo miembro no depende de θ y es fácil elegir la función $\varphi(x) \not\equiv 0$ que la reduce a cero.

§ 15. Familia exponencial

Supongamos que $\theta = (\theta_1, \ldots, \theta_k)$ es un parámetro k-dimensional y que la densidad $f_\theta(x)$ es representable en la forma

$$f_\theta(x) = h(x)\exp\left\{\sum_{j=1}^{k} a_j(\theta)U_j(x) + V(\theta)\right\}, \tag{1}$$

donde todas las funciones que entran en el segundo miembro son finitas y medibles.

---

*) Por $\mathfrak{U}_0$ aquí es necesario entender la σ-álgebra completada por los conjuntos N, para los cuales $\mathbf{P}_\theta(N) = 0$ para todos θ.

Definición 1. Las familias de distribuciones $\{\mathbf{P}_\theta\}$, con densidad de este género, se llaman *familias exponenciales* y se designan con el símbolo $\mathscr{E}$.

Para hacer que la representación (1) sea, en la medida de lo posible, unívoca, supondremos que las *funciones* $a_0(\theta) \equiv 1, a_1(\theta), \ldots, a_k(\theta)$ *son linealmente independientes en* Θ.

Como veremos, las familias exponenciales ocupan un lugar especial entre las familias paramétricas de distribuciones, ya que para ellas muchas construcciones generales de la estadística matemática pueden ser realizadas en forma explícita.

A veces se llaman familias exponenciales las familias de distribuciones de tipo más particular *), cuando $a_j(\theta) = \theta_j$.

A las familias exponenciales pertenecen, por ejemplo, las familias de distribuciones $\{\Phi_{\alpha,\sigma^2}\}$, $\{\Pi_\lambda\}\{\mathbf{B}_p\}, \{\Gamma_{\alpha,\lambda}\}$ y una serie de otras.

Ejemplo 1. Examinemos la distribución $\Gamma_{\alpha,\lambda}$. Su densidad $\gamma_{\alpha,\lambda}(x)$ se puede representar en la forma

$$\gamma_{\alpha,\lambda}(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x} = x^{-1} \exp\left\{\lambda \ln x - \alpha x + \ln \frac{\alpha^\lambda}{\Gamma(\lambda)}\right\}, \quad x > 0,$$

así que aquí se puede poner

$$h(x) = \begin{cases} x^{-1}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

$$U_1(x) = \ln x, \quad U_2(x) = x, \quad V(\alpha, \lambda) = \ln \frac{\alpha^\lambda}{\Gamma(\lambda)},$$

$$a_1(\alpha, \lambda) = \lambda, \quad a_2(\alpha, \lambda) = -\alpha. \quad \triangleleft$$

La función de verosimilitud para $X \in \mathbf{P} \in \mathscr{E}$ es igual a

$$f_\theta(X) = \exp\{(a(\theta), S) + nV(\theta)\} \prod_{i=1}^{n} h(x_i),$$

donde

$$a(\theta) = (a_1(\theta), \ldots, a_k(\theta)), \quad S = (S_1, \ldots, S_k),$$

$$S_j = S_j(X) = \sum_{i=1}^{n} U_j(x_i),$$

(a, S) es el producto escalar. De aquí y del teorema 12.1 resulta que S es una función suficiente para θ. Demostremos que S es una estadística suficiente mínima.

*) En realidad, esto es lo mismo; llegaremos a una forma particular si realizamos la transformación biunívoca $\gamma = \gamma(\theta)$, $\gamma = \gamma_1, \ldots, \gamma_k$) sobre el parámetro θ, poniendo $\gamma_j = a_j(\theta)$.

Como las funciones $a_j(\theta)$, $U_j(x)$, $V(\theta)$ son finitas, la exponencial en (1) es siempre positiva. Esto significa que en calidad de distribución $\mathbf{Q}$ en el teorema 13.1 (con la que todas las $\mathbf{P}_\theta$ son absolutamente continuas respecto a $\mathbf{P}_Q = \int \mathbf{P}_t \mathbf{Q}(dt)$) se puede tomar la distribución concentrada en cualquier punto fijado θ^0. Por eso, del teorema 13.1 se deduce que la σ-álgebra de $\mathfrak{U}_0$ engendrada por la función

$$r(X, \theta) = \frac{f_\theta(X)}{f_{\theta^0}(X)} = \exp\{(a(\theta) - a(\theta^0), S) + n(V(\theta) - V(\theta^0))\}$$

es la σ-álgebra suficiente mínima.

Teorema 1. *La estadística S es una estadística suficiente mínima.*

Demostración. De la independencia lineal de las funciones $1, a_1(\theta), \ldots, a_k(\theta)$ en Θ se deduce la independencia lineal $a_1(\theta) - a_1(\theta^0), \ldots, a_k(\theta) - a_k(\theta^0)$. Esto significa que en Θ hay k puntos $\theta^1, \ldots, \theta^k$ tales que los valores $a_{ij} = a_i(\theta^j) - a_i(\theta^0)$ forman una matriz A cuya determinante se distingue del cero. Esto significa, a su vez, que las ecuaciones $(a(\theta^j) - a(\theta^0), S) = \ln r(X, \theta^j) - n(V(\theta^j) - V(\theta^0))$, $j = 1, \ldots, k$, son solubles unívocamente respecto a S y, por lo tanto, $\sigma(S) \subset \sigma(r(X, \theta_j); j = 1, \ldots, k) \subset \mathfrak{U}_0$. ◁

En el ejemplo 1 hemos examinado la distribución Γ y establecimos que para ésta es válida la representación (1) cuando $\theta = (\alpha, \lambda)$ con las funciones

$$U_1(x) = \ln x, \quad U_2(x) = x,$$
$$a_1(\alpha, \lambda) = \lambda, \quad a_2(\alpha, \lambda) = -\alpha.$$

Es evidente que las condiciones del teorema 1 se han cumplido y que la estadística $S = \left(\sum \ln x_i, \sum x_i\right)$ o bien, que es lo mismo, la estadística $(\Pi x_i, \Sigma x_i)$ es una estadística suficiente mínima.

Si reforzamos un poco las condiciones del teorema 1, entonces la estadística S será una estadística suficiente completa (en este caso la minimización de S se podría obtener como consecuencia de la plenitud).

Teorema 2. *Sea $X \in \mathbf{P} \in \mathfrak{E}$ Si la función a y el conjunto Θ son tales que $a(\theta)$ traza un paralelepípedo k-dimensional cuando θ recorre Θ, entonces S es una estadística suficiente completa.*

Es evidente que las condiciones del teorema respecto al paralelepípedo se cumplirán si el conjunto Θ es "sólido", es decir, si contiene los puntos interiores (y junto con ellos también las esferas en R^k, de radio bastante pequeño) y si en el entorno de cualquier punto "sólido" θ^0, las funciones $a_j(\theta)$ son linealmente independientes y suaves. Entonces la transformación $a = a(\theta)$ transfiere el entorno del punto θ^0 al conjunto sólido.

Es evidente que el ejemplo 1, con la distribución Γ, satisface las condiciones del teorema 2, ya que la estadística $(\Pi x_i, \Sigma x_i)$ es completa.

De un modo igualmente sencillo, el lector puede comprobar que para la distribución normal Φ_{α,σ^2}, la estadística $\left(\sum x_i, \sum x_i^2\right)$ también es una estadística suficiente completa.

Demostración del teorema 2. En nuestro caso las funciones $\psi(s, \theta)$ y $h(x)$ en el teorema de factorización de Neyman — Fisher son iguales a

$$\psi(s, \theta) = \exp\{(a(\theta), s) + nV(\theta)\},$$

$$h(x) = \prod_{i=1}^{n} h(x_i).$$

Examinemos en $(R^k, \mathfrak{B}^k)$ la medida que no depende de θ:

$$\nu(B) = \int_{S^{-1}(B)} h(x)\mu^n(dx),$$

donde $S^{-1}(B)$ es el conjunto de todos los x para los cuales $S(x) \in B$.

Destaquemos en forma de lemas, las dos siguientes afirmaciones auxiliares.

Lema 1. *La distribución* $\mathbf{G}_\theta(B) = \mathbf{P}_\theta(S(X) \in B)$ *de la estadística* S *es absolutamente continua respecto a* ν, *y en el punto* s *tiene una densidad igual a* $\psi(s, \theta)$.

La **demostración** se deduce de la igualdad

$$\mathbf{G}_\theta(B) = \int_{S(x)\in B} \psi S(x), \theta)h(x)\mu^n(dx) = \int_{s\in B} \psi(s, \theta)\nu(ds),$$

la cual es consecuencia de la sustitución de las variables. ◁

Lema 2. *Sean* $\mathbf{G}_1$ *y* $\mathbf{G}_2$ *dos medidas* σ*-finitas en* $(R^k, \mathfrak{B}^k)$. *En este caso, si* $\int e^{(a,u)}\mathbf{G}_1(du) = \int e^{(a,u)}\mathbf{G}_2(du)$ *existen para todos los valores de* a *de cierto paralelepípedo* I *en* R^k, *entonces* $\mathbf{G}_1 = \mathbf{G}_2$.

Demostración. Para simplificar los razonamientos examinemos el caso unidimensional $k = 1$ y supongamos que $I = \{x : |x| \leqslant \alpha\}$. Entonces

$$h_j(a) = \int e^{au}\mathbf{G}_j(du), \quad j = 1, 2,$$

son funciones analíticas cuando $|a| < \alpha$. Además, para todos $b \in R$ están definidas las funciones $h_j(z) = \int e^{(a+ib)u}\mathbf{G}_j(du)$ de la variable compleja $z = a + ib$. Naturalmente que $h_j(z)$ serán analíticas en la franja de $|a| < \alpha$, $-\infty < b < \infty$. Como $h_1(z) = h_2(z)$ en el segmento de la recta $b = 0$, $|a| < \alpha$, entonces $h_1(z) = h_2(z)$ para todas z de la franja indicada. Por lo tanto,

$$\int e^{ibu}\mathbf{G}_1(du) = \int e^{ibu}\mathbf{G}_2(du). \tag{2}$$

Señalemos que en vista de que $h_j(0) = \int \mathbf{G}_j(du) < \infty$, podemos considerar que $\mathbf{G}_j$ son medidas probabilísticas. Del teorema de la correspondencia biunívoca entre las funciones características y las distribuciones [11], así como de (2), resulta que $\mathbf{G}_1 = \mathbf{G}_2$.

Si el paralelepípedo I tiene la forma $\{x: |x - \alpha_0| \leqslant \alpha\}$, entonces conviene pasar a las medidas $\mathbf{G}_j^{\alpha}(du) = e^{\alpha_0 u}\mathbf{G}_j(du)$.

En el caso multidimensional $k > 1$, la demostración se realiza exactamente igual. ◁

Ahora podemos pasar directamente *a la demostración del teorema 2.* Debemos demostrar que si φ es una función medible en $(R^k, \mathfrak{B}^k)$ y existe

$$\int \varphi(s)\mathbf{G}_\theta(ds) = 0 \text{ para todos } \theta \in \Theta, \tag{3}$$

entonces $\varphi(s) = 0$ c.d. $[\mathscr{G}]$, $\mathscr{G} = \{\mathbf{G}_\theta\}_{\theta\in\Theta}$. Sea $\varphi = \varphi^+ - \varphi^-$, donde $\varphi^{\pm} \geqslant 0$. En este caso, de (3) se desprende $\int \varphi^+(s)\mathbf{G}_\theta(ds) = \int \varphi^-(s)\mathbf{G}_\theta(ds)$ o bien, en virtud del lema 1,

$$\int \varphi^+(s)\psi(s, \theta)\nu(ds) = \int \varphi^-(s)\psi(s, \theta)\nu(ds),$$
$$\int \varphi^+(s)e^{(s, a(\theta))}\nu(ds = \int \varphi^-(s)e^{(s, a(\theta))}\nu(ds).$$

Si formamos las medidas σ-finitas $\nu^{\pm}(ds) = \varphi^{\pm}(s)\nu(ds)$, obtendremos

$$\int e^{(s, a)}\nu^+(ds) = \int e^{(s, a)}\nu^-(ds)$$

para todos los valores de a de cierto paralelepípedo en R^k. Sólo nos queda hacer uso del lema 2. ◁

Corolario 1. *Si* $X \in \mathbf{P} \in \mathscr{E}$, $\theta^* \in K_b$ *y se cumplen las condiciones del teorema 2, la estimación* $\theta_S^* = \mathbf{M}(\theta^*/S)$ *es la estimación eficiente en* K_b.

§ 16. Desigualdad de Rao — Cramer y estimaciones *R*-eficientes

1. Desigualdad de Rao — Cramer y sus corolarios. Los resultados de los párrafos precedentes nos proporcionaron varios criterios de eficacia de las estimaciones. Sin embargo, estos criterios tenían, en cierto sentido, un carácter cualitativo. En este párrafo continuaremos el estudio de la cuestión acerca de las estimaciones eficientes, pero desde un punto de vista un poco diferente. Aclaremos, ante todo, cuál es el valor mínimo del error estándar que se puede obtener.

Al principio examinaremos el caso unidimensional cuando θ es un parámetro escalar. Con respecto al conjunto Θ, para precisar vamos a suponer que eso es un intervalo finito o infinito, cerrado o abierto.

Para responder a la pregunta planteada necesitaremos las condiciones de regularidad en $f_\theta(x)$. Sea, como antes,

$$l(x, \theta) = \ln f_\theta(x), \; L(X, \theta) = \sum_{i=1}^{n} l(x_i, \theta), \; a(\theta) = \mathbf{M}_\theta \theta^* = \theta + b(\theta).$$

Supongamos que se ha cumplido la condición (R). *Las funciones* $\sqrt{f_\theta(x)}$ *para c.t.*$[\mu]$ *valores de* x *son continuamente derivables respecto a* $\theta \in \Theta$, *y la integral*

$$I(\theta) \equiv \int \frac{(f'_\theta(x))^2}{f_\theta(x)} \mu(dx) = \mathbf{M}_\theta [l'(x_1, \theta)]^2 \tag{1}$$

existe y es positiva y continua según θ. (Aquí y en lo sucesivo, la tilde significa la derivación respecto a θ).

Con arreglo a la integral (1) es necesario señalar lo siguiente: Si x, junto con su entorno, no pertenece al portador $N_{P_\theta} = \{x : f_\theta(x) > 0\}$ de la distribución $\mathbf{P}_\theta$, entonces la función subintegral $(f'_\theta(x))^2 / f_\theta(x)$ se convierte en indeterminación de tipo 0/0. Convendremos en considerar esta razón igual a cero. Seguiremos esa misma regla en cuanto a la derivada $l'(x, \theta) = f'_\theta(x)/f_\theta(x)$, al integrarla. Podríamos no hacer estas restricciones si desde el principio eximinaramos las integrales de la forma de $\mathbf{M}_\theta \varphi(x_1, \theta)$ sólo en la región de N_{P_θ}.

La función $I(\theta)$ es conocida con el nombre de *información de Fisher* y desempeña un papel muy importante en la matemática estadística, además, en lo sucesivo tropezaremos repetidas veces con ella. Algunas propiedades de la función $I(\theta)$ se examinan en § 17.

Si el conjunto Θ es compacto, la *continuidad de* $I(\theta)$ *en las condiciones* (R) *es equivalente a la condición*

$$\sup_{\theta \in \Theta} \mathbf{M}_\theta ([l'(x_1, \theta)]^2; \; |l'(x_1, \theta)| > N) \to 0$$

cuando $N \to \infty$, la cual se puede llamar convergencia uniforme de la integral $I(\theta)$ (véase el Suplemento VI).

Tiene lugar la siguiente desigualdad para la varianza de las estimaciones θ^* con desplazamiento b.

Teorema 1 (desigualdad de Rao — Cramer). *Si* $\theta^* \in K_b$ *y si está cumplida la condición* (R) *y* $\mathbf{M}_\theta(\theta^*)^2 < c < \infty$, *entonces*

$$\mathbf{D}_\theta \theta^* \geqslant \frac{[1 + b'(\theta)]^2}{nI(\theta)} . \tag{2}$$

Si en dicha desigualdad se alcanza igualdad en cierto segmento $\theta \in [\theta_1, \theta_2] \subset \Theta$, *y* $\mathbf{D}_\theta \theta^* > 0$ *en ese segmento, entonces la función de verosimilitud*

$f_\theta(X)$ para $\theta \in [\theta_1, \theta_2]$ es representable en la forma

$$f_\theta(X) = \exp\{\theta^* A(\theta) + B(\theta)\} h(X), \tag{3}$$

donde $A(\theta)$, $B(\theta)$ no dependen de X.

Al contrario, si $\theta^* = \text{const}$, o si es válida la representación (3), *entonces en la desigualdad* (2) *se alcanza igualdad.*

Evidentemente, la condición (3) significa que la distribución en $\mathscr{X}^n$ con densidad $f_\theta(x)$ pertenece a la familia exponencial $\mathscr{E}$.

Corolario 1. *Si se cumplen las condiciones del teorema* 1,

$$\mathbf{M}_\theta(\theta^* - \theta)^2 \geqslant \frac{[1 + b'(\theta)]^2}{nI(\theta)} + b^2(\theta).$$

Para cualquier estimación no desplazada θ^*,

$$\mathbf{M}_\theta(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)}.$$

Así pues, en las clases K_b, el valor mínimo posible de las desviaciones estándar es distinto de cero y se define por los segundos miembros de las desigualdades escritas.

Observación 1. En cuanto a la condición $\mathbf{M}_\theta(\theta^*)^2 < c < \infty$ se puede notar que cuando $\mathbf{M}_\theta(\theta^*)^2 = \infty$ se cumple $\mathbf{D}_\theta\theta = \infty$ y la desigualdad (2) se vuelve trivial. En virtud de (2), la condición $\mathbf{D}_\theta\theta > 0$ se puede sustituir por $(1 + b'(\theta))^2 > 0$.

Observación 2. A la par con la condición (R) se pueden señalar algunas otras condiciones que aseguran la afirmación del teorema 1 y que se distinguen muy poco una de otra. Nos hemos detenido en aquellas de ellas que nos serán más cómodas en los párrafos posteriores. Las condiciones de tipo algo diferente se citarán en el § 22.

Necesitaremos una afirmación auxiliar.

Lema 1. *Supongamos que se ha cumplido la condición (R) y que $S = S(X)$ es cualquier estadística para la cual $\mathbf{M}_\theta S^2 < c < \infty$ cuando $\theta \in \Theta$. Entonces la función*

$$a_S(\theta) = \mathbf{M}_\theta S = \int S(x) f_\theta(x) \mu^n(dx) \tag{4}$$

es derivable respecto a θ, además

$$a_S'(\theta) = \int S(x) f_\theta'(x) \mu^n(dx) = \mathbf{M}_\theta S L'(X, \theta). \tag{5}$$

Esta afirmación tiene carácter técnico y su demostración dificultaría considerablemente las investigaciones. Por eso hemos pasado la demostración del lema 1 al Suplemento VI.

Demostración del teorema 1. Poniendo en (5) $S \equiv 1$, obtenemos $a_S(\theta) \equiv 1$,

$$\mathbf{M}_\theta L' = 0, \ \mathbf{M}_\theta a(\theta) L' = 0. \tag{6}$$

Volviendo a utilizar (5) para $S = \theta^*$ y (6), obtenemos

$$\mathbf{M}_\theta \theta^* L' = a'(\theta), \ \mathbf{M}_\theta(\theta^* - a(\theta))L' = a'(\theta). \tag{7}$$

Según la desigualdad de Cauchy — Buniakovski,

$$(a'(\theta))^2 \leqslant \mathbf{M}_\theta(\theta^* - a(\theta))^2 \mathbf{M}_\theta(L')^2 \tag{8}$$

o bien, que es lo mismo,

$$\mathbf{D}_\theta \theta^* \geqslant \frac{(1 + b'(\theta))^2}{\mathbf{M}_\theta(L')^2}. \tag{9}$$

Como las variables aleatorias $l_j = l'(x_j, \theta)$ son independientes, están igualmente distribuidas y tienen, en virtud de (6), una esperanza matemática nula, $\mathbf{M}_\theta l_j = 0$, entonces $\mathbf{M}_\theta l_i l_j = 0$ cuando $i \neq j$,

$$\mathbf{M}_\theta(L')^2 = \mathbf{M}_\theta\left(\sum_j l_j\right)^2 = \sum_{i,j} \mathbf{M}_\theta l_i l_j = n\mathbf{M}_\theta l_1^2 = nI(\theta).$$

Junto con (9) esto demuestra la desigualdad (2).

Demostremos ahora la segunda afirmación del teorema. Para simplificar la demostración consideraremos que Θ coincide con $[\theta_1, \theta_2]$ y que la medida μ está concentrada en la unión de los portadores de $\mathbf{P}_\theta$, $\theta \in \Theta$. El signo de igualdad en (2) (o en 8)) quiere decir que

$$\int (\theta^* - a(\theta)) f'_\theta(x) \mu^n(dx) = \left[\int (\theta^* - a(\theta))^2 f_\theta(x) \mu^n(dx) \int \frac{(f'_\theta(x))^2}{f_\theta(x)} \mu^n(dx)\right]^{1/2}$$

para todos $\theta \in \Theta$. En vista de que la primera integral en el segundo miembro es positiva, la igualdad escrita sólo será posible si

$$f'_\theta(x)/\sqrt{f_\theta(x)} = c(\theta)(\theta^* - a(\theta))\sqrt{f_\theta(x)} \ \text{ c.t.}[\mu^n]. \tag{10}$$

Designemos por A el conjunto de x para los que está cumplida (10) y $|\theta^*| < \infty$. Entonces $\mu(\overline{A}) = 0$ ($\overline{A}$ es el complemento a A). Anotamos $x \in A$. En virtud de la continuidad $f_\theta(x)$ en θ, tendremos $f_1(x) > 0$ en cierto intervalo $(t_1, t_2) \subset \Theta$, y en este intervalo, en virtud de (10),

$$L'(x, \theta) = c(\theta)(\theta^* - a(\theta)). \tag{11}$$

Señalemos ahora, que de (7), (11) y (2) resulta

$$a'(\theta) = \mathbf{M}_\theta(\theta^* - a(\theta))L' = c(\theta)\mathbf{D}_\theta\theta^*, \ \mathbf{D}_\theta\theta^* = \frac{(a'(\theta))^2}{nI(\theta)}, \tag{12}$$

$$|c(\theta)| = \sqrt{\frac{nI(\theta)}{\mathbf{D}_\theta\theta^*}},$$

así que $\mathbf{D}_\theta\theta^*$ es continua en θ junto con $a'(\theta)$, $I(\theta)$, y $|c(\theta)|$ junto con $a(\theta)$ están limitadas uniformemente en $[\theta_1, \theta_2]$. La derivada $L'(x, \theta)$ en (11) posee esa misma propiedad. Pero esto significa que $L(x, t)$ es finita y que $f_\theta(x) > 0$ en todas las partes de $\Theta = [\theta_1, \theta_2]$, así que (11) es válida para todos θ. Integrando (11) dentro de los límites de θ_1 y θ, obtendremos

$$L(x, \theta) = \theta^* \int_{\theta_1}^{\theta} c(t)dt - \int_{\theta_1}^{\theta} c(t)a(t)dt + L(x, \theta_1),$$

que es equivalente a (3) para $[\mu^n]$ c.t. x. Como la variación $f_\theta(x)$ en el conjunto de la μ^n-medida 0 no tiene importancia, (3) queda demostrada.

Examinemos ahora la última afirmación del teorema. Si $\theta^* = \text{const}$, entonces $b'(\theta) = -1$ y ambos miembros de la desigualdad (2) se anulan. Supongamos que se ha cumplido (3). Entonces, derivando la función $L(X, \theta)$ respecto a θ, obtendremos

$$L'(X, \theta) = \theta^* A'(\theta) + B'(\theta).$$

De (7) se deduce que $a(\theta)A'(\theta) + B'(\theta) = 0$. Por eso

$$L'(X, \theta) = A'(\theta)(\theta^* - a(\theta))$$

y, por consiguiente (véase (10)), en (2) se alcanza la igualdad. ◁

En lo sucesivo excluiremos de las investigaciones el caso trivial $\theta^* = \text{const}$ y supondremos que $\mathbf{D}_\theta\theta^* > 0$ en todas las partes de Θ. Entonces es válido el

Corolario 2. *Si se cumplen las condiciones (R), para alcanzar la frontera inferior en la desigualdad de Rao — Cramer es necesario y suficiente que la estimación θ^* sea suficiente y que la función $\psi(\theta^*, \theta)$ en la igualdad de factorización tenga la forma*

$$\psi(\theta^*, \theta) = \exp\{\theta^* A(\theta) + B(\theta)\},$$

donde $A(\theta)$ y $B(\theta)$ son funciones derivables.

Corolario 3. *Si se cumplen las condiciones (R), $\theta^* \in K_b$, y en la desigualdad de Rao — Cramer se alcanza igualdad, entonces θ^* es una estimación eficiente en K_b.*

Esta afirmación se deduce de la representación

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \mathbf{D}_\theta\theta^* + b^2(\theta).$$

Señalemos que, hablando en general, lo contrario no es cierto: la estimación

puede ser eficiente en K_b, pero la frontera inferior $\dfrac{(1 + b'(\theta))^2}{nI(\theta)}$ para la varianza puede no alcanzarse.

Ejemplo 1. Sea $X \in \Gamma_{\alpha,1}$. Aquí $f_\alpha(X) = \alpha^n e^{-\alpha n\bar{x}}$. Las condiciones (R) en la región $\Theta \subseteq \{\alpha \geqslant \delta > 0\}$ están cumplidas. Es evidente que $S = n\bar{x}$ es una estadística suficiente completa. Por eso la estimación $\alpha^* = \bar{x}^{-1} = \mathbf{M}_\alpha(\bar{x}^{-1}/S)$ es eficiente en la clase K_b con un desplazamiento $b(\alpha) = \mathbf{M}_\alpha x^{-1} - \alpha$.

Notemos ahora que $S \in \Gamma_{\alpha,n}$, así que cuando $n > 1$ (véase el § 2), $\mathbf{M}_\alpha \bar{x}^{-1} = n\mathbf{M}_\alpha S^{-1} = \dfrac{n}{n-1}\,\alpha$.

Ahora bien, la estimación $\alpha^{**} = \dfrac{n-1}{n\bar{x}} = \alpha^*\left(1 - \dfrac{1}{n}\right)$ no estará desplazada cuando $n > 1$. Análogamente, cuando $n > 2$ hallamos (véase el § 2 y también ejemplo 4.1)

$$\mathbf{M}_\alpha(\alpha^{**})^2 = (n-1)^2\mathbf{M}_\alpha S^{-2} = \frac{n-1}{n-2}\,\alpha^2,$$

$$\mathbf{D}_\alpha\alpha^{**} = \alpha^2\left[\frac{n-1}{n-2} - 1\right] = \frac{\alpha^2}{n-2}.$$

Así pues, cuando $n > 2$, la estimación α^{**} es eficiente. Sin embargo, el criterio (3) no se ha cumplido, ya que

$$f_\alpha(X) = \alpha^{-n} e^{-\alpha(n-1)/\alpha^{**}}.$$

Por consiguiente, en la desigualdad de Rao — Cramer no se alcanza la frontera inferior. De esto también podemos convencernos directamente. En efecto, aquí $l(x, \alpha) = \ln \alpha - \alpha x$, $l'(x, \alpha) = 1/\alpha - x$ e

$$I(\alpha) = \mathbf{M}_\alpha[l'(x_1, \alpha)]^2 = \mathbf{M}_\alpha\left(\frac{1}{\alpha} - x_1\right)^2 = \frac{1}{\alpha^2} - \frac{2}{\alpha^2} + \frac{2}{\alpha^2} = \frac{1}{\alpha^2}.$$

Por lo tanto, cuando $n > 2$,

$$\frac{1}{nI(\theta)} = \frac{\alpha^2}{n} < \frac{\alpha^2}{n-2} = \mathbf{D}_\alpha\alpha^{**}.$$

Ahora bien, el logro de la frontera inferior en (2) es una exigencia más severa que el logro de la eficacia.

2. Estimaciones *R*-eficientes y asintóticamente *R*-eficientes. Supongamos que se han cumplido las condiciones (R). En este caso, el logro de la frontera inferior (exacto o asintótico) para la varianza en la desigualdad de Rao — Cramer puede ser un índice muy importante de la calidad de las estimaciones, íntimamente ligado al concepto de eficacia.

Definición 1. La estimación $\theta^* \in K_b$, para la cual

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \frac{(1 + b'(\theta))^2}{nI(\theta)} + b^2(\theta),$$

se llama *R-eficiente* (o *regularmente eficiente*) *en la clase* K_b.

La estimación R-eficiente en la clase K_0 de las estimaciones no desplazadas se denomina simplemente *R-eficiente.*

La estimación θ^* se denomina *asintóticamente R-eficiente* (a.R-e.), si

$$\mathbf{M}_\theta(\theta^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

Vemos que a diferencia de las definiciones del § 8, que tenían un carácter más cualitativo, las definiciones de *R*-eficacia se basan en la comparación con los valores numéricos conocidos, relacionados principalmente con la información de Fisher, mejor dicho, con la cantidad $(nI(\theta))^{-1}$.

Para la *R*-eficacia de θ^* es necesario y suficiente el cumplimiento de (3).

De lo dicho más arriba se deduce que las estimaciones *R*-eficientes son eficientes, pero no al revés; las estimaciones *R*-eficientes simplemente existen con menos frecuencia, lo cual no es un defecto de las estimaciones, sino de la frontera inferior en la desigualdad de Rao — Cramer.

En los actuales manuales de estadística matemática, las estimaciones *R*-eficientes se llaman simplemente eficientes. No obstante, creemos que es más natural conservar el término «eficacia» para las mejores estimaciones en un sentido más amplio (véase la definición 8.1).

Teorema 2. *Si se han cumplido las condiciones* (*R*) *y existe la estimación R-eficiente, entonces esta última coincide con la estimación de verosimilitud máxima.*

Demostración. Ya hemos visto que el cumplimiento de (3) conduce a la igualdad (véase (11))

$$L'(X, \theta) = (\theta^* - \theta)c(\theta).$$

Además, como $b(\theta) = 0$, de (12) resulta

$$c(\theta) = 1/\mathbf{D}_\theta\theta^* = nI(\theta) > 0$$

para cualesquier $\theta \in \Theta$. Esto quiere decir que $L'(X, \theta) < 0$ cuando $\theta > \theta^*$, y que $L'(X, \theta) > 0$ cuando $\theta < \theta^*$. Por consiguiente, cuando $\theta = \theta^*$ se alcanza el máximo $L(X, \theta)$. ◁

El ejemplo 1 citado más arriba muestra que, a diferencia de las estimaciones *R*-eficientes, las estimaciones eficientes pueden no coincidir con las e.v.m. En este ejemplo, la e.v.m. es $(\overline{x})^{-1}$, mientras que la estimación eficien-

te es igual a $\frac{n-1}{n}(\overline{\chi})^{-1}$. Estas dos estimaciones son, evidentemente, las estimaciones a.*R*-e.

Examinemos la clase $\tilde{K}_0$ de las estimaciones θ^*, para las cuales, cuando $n \to \infty$,

$$|b(\theta)| \leqslant \varepsilon(\theta, n)/\sqrt{n}, \ |b'(\theta)| \leqslant \varepsilon(\theta, n),$$
$$\mathbf{M}_\theta(\theta^*)^2 < c < \infty$$

para cierta función $\varepsilon(\theta, n) = o(1)$ cuando $n \to \infty$ y cuando cada $\theta \in \Theta$.

Cada una de estas clases es notable por el hecho de que para ella la frontera inferior en la desigualdad de Rao — Cramer tiene la forma $(1 + o(1))/[nI(\theta)]$. En el § 20 veremos que en una serie de casos, al hallar las estimaciones asintóticamente óptimas, es posible limitarse al estudio de las estimaciones θ^* de tales clases.

Teorema 3. *Supongamos que se han cumplido las condiciones* (*R*). *Entonces, cualquier estimación a.R-e. de* $\tilde{K}_0$ *es la estimación. a.e. en* $\tilde{K}_0$.

La demostración del teorema es evidente: si θ_1^* es la estimación a.*R*-e., entonces

$$\mathbf{M}_\theta(\theta_1^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

Además, como ya hemos señalado, según la desigualdad de Rao — Cramer, para todos $\theta^* \in \tilde{K}_0$,

$$\lim_{n\to\infty} \inf \mathbf{M}_\theta n(\theta^* - \theta)^2 \geqslant I^{-1}(\theta) = \lim_{n\to\infty} \mathbf{M}_\theta n(\theta_1^* - \theta)^2. \ \triangleleft$$

También está claro que si existe la estimación a.*R*-e., cualquier estimación a.e. en $\tilde{K}_0$ será la estimación a.*R*-e.

Más tarde (véase el § 25) veremos que con ciertas suposiciones adicionales, las estimaciones a.*R*-e. existen siempre y, por consiguiente, la afirmación del teorema 3 también es válida en dirección inversa: la estimación a.e. en $\tilde{K}_0$ es la estimación a.*R*-e. o sea, para ella $\mathbf{M}_\alpha(\theta^* - \theta)^2 \sim |nI(\theta)|^{-1}$.

Teorema 4. *Supongamos que se han cumplido las condiciones* (*R*). *Si* θ_1^*, θ_2^* *pertenecen a* $\tilde{K}_0$ *y son las estimaciones a.R-e., ellas son asintóticamente equivalentes en el sentido siguiente:*

$$\sqrt{n}(\theta_1^* - \theta_2^*) \underset{P}{\to} 0.$$

La demostración de esta afirmación se efectúa exactamente igual que en el teorema 8.2. Como $\theta^* = (\theta_1^* + \theta_2^*)/2 \in \tilde{K}_0$, entonces, basándonos en (8.11) y en la igualdad de Rao — Cramer, obtenemos

$$\lim_{n\to\infty} \sup \mathbf{M}_\theta n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \quad \triangleleft$$

Ejemplo 2. La estimación $\alpha^* = \bar{x}$ del valor medio α de la población normal Φ_{α,σ^2} para σ^2 conocida es la estimación R-eficiente. Es fácil convencerse de esto, comprobando, por ejemplo, la condición (3). Otra posibilidad consiste en comparar $\mathbf{D}_\alpha\alpha^* = \sigma^2/n$ con el valor mínimo posible $(nI(\alpha))^{-1}$ de las varianzas de las estimaciones no desplazadas. En nuestro caso,

$$l(x, \alpha) = -\ln \sqrt{2\pi}\,\sigma - (x-\alpha)^2/(2\sigma^2),$$
$$l'(x, \alpha) = (x-\alpha)/\sigma^2,$$
$$I(\alpha) = \mathbf{M}_\alpha[l'(x_1, \alpha)]^2 = \mathbf{M}_\alpha(x_1-\alpha)^2/\sigma^4 = 1/\sigma^2,$$

así que $\mathbf{D}_\alpha\alpha^* = (nI(\alpha))^{-1} = \sigma^2/n$.

Ejemplo 3. Examinemos la estimación $\theta^* = S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\alpha)^2$ del parámetro $\theta = \sigma^2$ de la población normal con α conocido. No es difícil calcular que $\mathbf{D}_\theta\theta^* = \mathbf{M}_\theta(\theta^*-\sigma^2)^2 = 2\sigma^4/n$. Por otro lado, aquí

$$l'(x_1, \theta) = -\frac{1}{2\theta} + \frac{(x_1-\alpha)^2}{2\theta^2},$$

$$I(\theta) = \mathbf{M}_\theta[l'(x_1, \theta)]^2 = \frac{1}{4\theta^4}\,\mathbf{M}_\theta[(x_1-\alpha)^2 - \theta]^2 = \frac{1}{2\theta^2} = \frac{1}{2\sigma^4}.$$

Ahora bien, aquí también $\mathbf{D}_\theta\theta^* = (nI(\theta))^{-1}$, y la estimación $\theta^* = S_1^2$ es R-eficiente.

La varianza de la estimación no desplazada $S_0^2 = \frac{1}{n-1}\sum(x_i-\bar{x})^2$ es igual a $\frac{2\sigma^4}{n-1}$, así que la misma no es R-eficiente o simplemente no es la estimación eficiente de σ^2. Al mismo tiempo es evidente que S_0^2 es la estimación a.R-e.

Si en calidad de parámetro desconocido estimamos no σ^2 sino $\theta = \sigma$, entonces no obtendremos la estimación R-eficiente. Sin embargo, la estimación no desplazada de σ será la estimación

$$\sigma^* = \sqrt{\frac{n}{2}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\,S, \text{ ya que}$$

$$\mathbf{M}_\sigma S = \frac{\sigma}{\sqrt{n}}\,\mathbf{M}_\sigma\sqrt{\frac{1}{\sigma^2}\sum(x_i-\alpha)^2},$$

donde $\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum(x_i-\alpha)^2$ tiene la distribución $\mathbf{H}_n = \Gamma_{1/2,n/2}$, por eso (véase el § 2)

$$\mathbf{M}_\sigma S = \frac{\sigma\sqrt{2}}{\sqrt{n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \quad \mathbf{M}_\sigma \sigma^* = \sigma.$$

Como S es la estimación suficiente completa y mínima, σ^* es la estimación eficiente. Con ayuda de la fórmula de Stirling no es difícil convencerse de que $\sigma^* = S(1 + O(1/n))$.

Comparemos ahora la magnitud $\mathbf{D}_\sigma \sigma^*$ con la frontera inferior $(nI(\sigma))^{-1}$. Tenemos

$$\mathbf{D}_\sigma \sigma^* = \mathbf{M}_\sigma \left(\sqrt{\frac{n}{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} S - \sigma \right)^2 = \sigma^2 \left[\frac{n}{2} \frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n+1}{2}\right)} - 1 \right]. \tag{13}$$

Por otro lado, aquí

$$l'(x, \sigma) = -\frac{1}{\sigma} + \frac{(x-\alpha)^2}{\sigma^3},$$

$$I(\sigma) = \mathbf{M}_\sigma [l'(x_1, \sigma)]^2 = \frac{1}{\sigma^6} \mathbf{M}_\sigma [(x_1 - \alpha)^2 - \sigma^2]^2 = \frac{2}{\sigma^2},$$

así que $(nI(\sigma))^{-1} = \sigma^2/(2n)$. Pero este valor se distingue de (13). Su relación, por ejemplo para $n = 3$, es igual a 0,936. Ahora bien, aquí no hay estimaciones R-eficientes. Cuando $n \to \infty$ el coeficiente de σ^2 en (13) se comporta asintóticamente como $\frac{1}{2n} + O\left(\frac{1}{n^2}\right)$, así que σ^* es la estimación a.R-e.

3. **Desigualdad de Rao — Cramer en el caso multidimensional.** En este apartado $\theta = (\theta_1, \ldots, \theta_k)$ es el vector k-dimensional, al igual que también la estimación $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$. Como antes, pongamos

$$a(\theta) = \mathbf{M}_\theta \theta^* = \theta + b(\theta), \quad b(\theta) = (b_1(\theta), \ldots, b_k(\theta))$$

y examinemos las clases K_b de las estimaciones con un desplazamiento registrado $b(\theta)$.

La generalización de las condiciones (R) para el caso multidimensional tendrá el aspecto siguiente. Designemos

$$l(x, \theta) = \log f_\theta(x), \quad l_i(x, \theta) = \frac{\partial}{\partial \theta_i} l(x, \theta),$$

$$I_{ij}(\theta) = \mathbf{M}_\theta l_i(x_1, \theta) l_j(x_1, \theta)$$

y supongamos que se ha cumplido la condición

(R). *Las funciones* $\sqrt{f_\theta(x)}$ *son derivables continuamente respecto a* θ_j *para c.t.* $[\mu]$ *valores de* x. *La matriz*

$$I(\theta) = \|I_{ij}(\theta)\|,$$

$$I_{ij}(\theta) = \int l_i(x; \theta) l_j(x; \theta) f_\theta(x) \mu(dx)$$

es continua en θ **), y su determinante* $|I(\theta)|$ *es distinto del cero.*

Como $I(\theta)$ es la matriz de segundos momentos $\mathbf{M}_\theta l_i l_j$ de las variables aleatorias $l_i = l_i(x_1, \theta)$, ella será una matriz definida positivamente, ya que para cualquier vector $\alpha = (\alpha_1, \ldots, \alpha_k) \neq 0$ se cumple

$$\sum \alpha_i \alpha_j \mathbf{M}_\theta l_i l_j = \mathbf{M}_\theta \left(\sum \alpha_i l_i\right)^2 \geqslant 0,$$

donde la igualdad a cero se exluye por la condición $|I(\theta)| \neq 0$.

Como antes, por desigualdad entre las matrices $\sigma_1^2 \geqslant \sigma_2^2$ entenderemos la desigualdad $\alpha\sigma_1^2\alpha^T \geqslant \alpha\sigma_2^2\alpha^T$ para cualquier vector fila $\alpha = (\alpha_1, \ldots, \alpha_k) \neq 0$. Esto equivale, evidentemente, al hecho de que la matriz $\sigma_1^2 - \sigma_2^2$ está definida de forma no negativa. La desigualdad estricta corresponderá a la definición positiva, así que, por ejemplo, $I(\theta) > 0$.

Teorema 1A. *Si* $\theta^* \in K_b$ *y si se cumple la condición* (R)*, entonces para la matriz de segundos momentos* $\sigma^2 = \|\sigma_{ij}\| = \mathbf{M}_\theta(\theta^* - a(\theta))^T(\theta^* - a(\theta))$ *de cualquier estimación* θ^* *del vector fila* θ *es válida la desigualdad*

$$\sigma^2 \geqslant \frac{1}{n}(E + D(\theta))I^{-1}(\theta)(E + D(\theta))^T, \tag{14}$$

donde E *es la matriz unidad,* $D(\theta) = \|b_{ij}(\theta)\|$, $b_{ij}(\theta) = \dfrac{\partial b_i(\theta)}{\partial \theta_j}$.

Sea $|\sigma|^2 > 0$ *(o bien* $|E + D(\theta)| > 0$*) para todos* θ*. En este caso el signo de igualdad en* (14) *se alcanza si y sólo si la distribución de la muestra pertenece a una familia exponencial de tipo especial, o sea, cuando para ciertas funciones escalares* $B(\theta)$ *y* $h(X)$ *se cumple*

$$f_\theta(X) = \exp\{(\theta^*, A(\theta) + B(\theta)\}h(X), \tag{15}$$

donde el vector $A(\theta) = (A_1(\theta), \ldots, A_k(\theta))$ *tiene una matriz de derivadas igual a*

$$\|A_{ij}\| = \left\| \frac{\partial A_i(\theta)}{\partial \theta_j} \right\| = n[(E + D(\theta))^{-1}]^T I(\theta).$$

Es evidente que para las estimaciones no desplazadas θ^*,

$$\sigma^2 \geqslant (nI(\theta))^{-1}$$

y la igualdad es posible únicamente cuando se cumple (15)*, donde* $|A_{ij}| = nI(\theta)$.

Ahora bien, si logramos hallar la estimación no desplazada θ^* con una matriz de segundos momentos $[nI(\theta)]^{-1}$, ella será una estimación eficiente.

---

*) Para esto es suficiente exigir la convergencia uniforme de $I_{it}(0)$ (véase el Suplemento VI).

En el caso multidimensional conservan su validez todas las observaciones hechas con arreglo a la desigualdad unidimensional de Rao — Cramer, así como la definición de R-eficacia, en las que deben introducirse tan sólo las modificaciones evidentes relacionadas con la dimensión de θ.

En particular, llamaremas estimaciones a.R-e. las estimaciones θ^* para las cuales

$$\mathbf{M}_\theta(\theta^* - \theta)^T(\theta^* - \theta) = \sigma^2 + b^T(\theta)b(\theta) = (nI(\theta))^{-1} + o(1/n).$$

Aquí el análogo del teorema 2 tendrá el aspecto siguiente.

Teorema 2A. *Supongamos que se cumplen las condiciones* (R). *Si* θ^* *es la estimación R-eficiente, entonces ésta es la estimación de verosimilitud máxima.*

Demostración. Para demostrar que la estimación R-eficiente constituye el único punto del máximo, es suficiente convencerse que $L'(X, \theta^*) = 0$ y que cuando $\theta = \theta^* + u$, $u \neq 0$,

$$(\operatorname{grad} L(X, \theta), u) = (L'(X, \theta), \theta - \theta^*) < 0.$$

Pero en el caso de existencia de la estimación R-eficiente, se cumple (véase (20))

$$L'(X, \theta) = (\theta^* - \theta)nI(\theta),$$

de donde se desprenden inmediatamente las relaciones requeridas. La segunda se deduce del hecho de que

$$(L', u) = -unI(\theta)u^T,$$

donde $uI(\theta)u^T$ es la forma cuadrática definida positivamente. ◁

Ejemplo 4. Examinemos una familia biparamétrica de distribuciones normales Φ_{α,σ^2}. La misma pertenece a una familia exponencial, ya que (aquí $\theta = (\theta_1, \theta_2)$, $\theta_1 = \alpha$, $\theta_2 = \sigma^2$)

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} - \ln\sigma\right\}.$$

La estimación $\theta^* = (\theta_1^*, \theta_2^*)$, donde $\theta_1^* = \bar{x}$, $\theta_2^* = S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - \bar{x}^2\right)$ es eficiente, puesto que pertenece a K_0, y la estadística $\left(\sum x_i, \sum x_i^2\right)$, como hemos visto en el § 15, es la estadística suficiente completa (véase el teorema 14.4).

Señalemos que

$$\mathbf{M}_\theta(\theta^* - \theta)^T(\theta^* - \theta) = \sigma^2 + b^T(\theta)b(\theta).$$

Demostración del teorema 1A. Designemos

$$L_j = L_j(X, \theta) = \sum_{i=1}^{n} l_j(x_i, \theta), \; L' = L'(X, \theta) = (L_1, \ldots, L_k).$$

Entonces, de un modo completamente análogo al caso unidimensional, establecemos que son válidas las igualdades

$$\mathbf{M}_\theta l_j(x_1, \theta) = 0, \; \mathbf{M}_\theta l_i L_j(X, \theta) = 1 + b_{ij}(\theta)$$

en las cuales $b_{ij}(\theta)$ son continuas o bien, que es lo mismo, las igualdades

$$\mathbf{M}_\theta L' = 0, \tag{16}$$

$$\mathbf{M}_\theta(\theta^*)^T L' = E + D(\theta) \tag{17}$$

en las que la matriz $D(\theta)$ es continua. De aquí obtenemos

$$\mathbf{M}_\theta(\theta^* - a(\theta))^T L' = E + D(\theta). \tag{18}$$

Demostremos ahora la desigualdad siguiente (variante matricial de la desigualdad de Cauchy — Buniakovski).

Lema 2. *Supongamos que ξ y η son matrices de igual dimensión (no obligatoriamente cuadradas) con elementos aleatorios, y que la matriz $\mathbf{M}\eta\eta^T$ tiene inversa. Entonces*

$$\mathbf{M}\xi\xi^T \geqslant \mathbf{M}\xi\eta^T(\mathbf{M}\eta\eta^T)^{-1}\mathbf{M}\eta\xi^T. \tag{19}$$

En este caso la igualdad es posible únicamente cuando $\xi = z\eta$, $z = \mathbf{M}\xi\eta^T(\mathbf{M}\eta\eta^T)^{-1}$.

Demostración. En vista de que para cualquier matriz A es válida la desigualdad $AA^T \geqslant 0$ (AA^T está definida no negativamente), entonces

$$0 \leqslant \mathbf{M}(\xi - z\eta)(\xi - z\eta)^T = \mathbf{M}\xi\xi^T - z\mathbf{M}\eta\xi^T - \mathbf{M}\xi\eta^T z^T + z\mathbf{M}\eta\eta^T z^T.$$

Poniendo $z = \mathbf{M}\xi\eta^T(\mathbf{M}\eta\eta^T)^{-1}$, obtenemos la desigualdad requerida.

La afirmación con respecto a las condiciones de la igualdad en (19) es evidente. ◁

Volvamos a la demostración del teorema 1A. Pongamos, en (19), $\xi = (\theta^* - a(\theta))^T$, $\eta = (L')^T$. Entonces

$$\mathbf{M}_\theta\xi\xi^T = \mathbf{M}_\theta(\theta^* - a(\theta))^T(\theta^* - a(\theta) = \sigma^2.$$

De (16) y de la desigualdad de x_i obtenemos

$$\mathbf{M}_\theta\eta\eta^T = \mathbf{M}_\theta(L')^T L' = nI(\theta).$$

Por último, de (18) hallamos

$$\mathbf{M}_\theta \xi \eta^T = \mathbf{M}_\theta(\theta^* - a(\theta))^T L' = E + D(\theta).$$

La desigualdad (14) queda demostrada.

La desigualdad en (14) es posible en virtud del lema 2, si sólo para los puntos (x, θ), tales que $f_\theta(x) > 0$, es válida

$$(\theta^* - a(\theta))^T = (E + D(\theta))(nI(\theta)^{-1}(L')^T$$

o, que es lo mismo,

$$L' = (\theta^* - a(\theta))n[(E + D(\theta))^{-1}]^T I(\theta). \tag{20}$$

Nótese ahora que de la desigualdad en (14) resulta

$$|E + D(\theta)|^2 = n|\sigma^2| \cdot |I(\theta)|,$$

y la separación del determinante $|\sigma^2|$ de 0 quiere decir lo mismo para $|E + D(\theta)|$ y significa la existencia de la matriz inversa $(E + D(\theta))^{-1}$ uniformemente limitada. Por eso la derivada L' en (20) será limitada, y $f_\theta(x) > 0$ en todas partes de Θ y la misma igualdad (20) será válida en todas partes de Θ. Si ahora s es cualquier camino que une los puntos θ_1 y θ en la región Θ, entonces

$$L(X, \theta) = \int_s (L', ds) + L(X, \theta_0),$$

donde ds significa el elemento vectorial del camino s; $((L', ds) = (L', s'(l))dl$ es el incremento $L(X, \theta)$ en dicho camino; y l, la «longitud» del camino recorrido. Por consiguiente, en virtud de (20),

$$L(X, \theta) = \theta^* A(\theta) + B(\theta) + H(X), \tag{21}$$

donde $B(\theta)$ y $H(X)$ son funciones escalares; $A(\theta) = (A_1(\theta), \ldots, A_k(\theta))$ es un vector que depende exclusivamente de sus argumentos. Esto significa la validez de (15).

Si se cumple (21), entonces

$$L' = \theta^* |A_{ij}| + B'(\theta),$$

donde, en virtud de la igualdad $\mathbf{M}_\theta L' = 0$, es válida

$$B'(\theta) = -a(\theta)|A_{ij}|.$$

Multiplicando ambos miembros de la igualdad $L' = (\theta^* - a(\theta))|A_{ij}|$, a la izquierda en $(\theta^* - a(\theta))^T$, obtenemos, en virtud de (18), que para el cumplimiento de la condición (20), que significa la igualdad en (14), debe cumplirse

$$|A_{ij}| = n[(E + D(\theta))^{-1}]^T I(\theta). \quad \triangleleft$$

En el caso multidimensional conservan su validez todas las observaciones hechas con arreglo a la desigualdad unidimensional de Rao — Cramer, así como la definición de R-eficacia, en las que deben introducirse tan sólo las modificaciones evidentes relacionadas con la dimensión de θ.

En particular, llamaremas estimaciones a.R-e. las estimaciones θ^* para las cuales

$$\mathbf{M}_\theta(\theta^* - \theta)^T(\theta^* - \theta) = \sigma^2 + b^T(\theta)b(\theta) = (nI(\theta))^{-1} + o(1/n).$$

Aquí el análogo del teorema 2 tendrá el aspecto siguiente.

Teorema 2A. *Supongamos que se cumplen las condiciones (R). Si θ^* es la estimación R-eficiente, entonces ésta es la estimación de verosimilitud máxima.*

Demostración. Para demostrar que la estimación R-eficiente constituye el único punto del máximo, es suficiente convencerse que $L'(X, \theta^*) = 0$ y que cuando $\theta = \theta^* + u$, $u \neq 0$,

$$(\operatorname{grad} L(X, \theta), u) = (L'(X, \theta), \theta - \theta^*) < 0.$$

Pero en el caso de existencia de la estimación R-eficiente, se cumple (véase (20))

$$L'(X, \theta) = (\theta^* - \theta)nI(\theta),$$

de donde se desprenden inmediatamente las relaciones requeridas. La segunda se deduce del hecho de que

$$(L', u) = -unI(\theta)u^T,$$

donde $uI(\theta)u^T$ es la forma cuadrática definida positivamente. ◁

Ejemplo 4. Examinemos una familia biparamétrica de distribuciones normales Φ_{α,σ^2}. La misma pertenece a una familia exponencial, ya que (aquí $\theta = (\theta_1, \theta_2)$, $\theta_1 = \alpha$, $\theta_2 = \sigma^2$)

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\alpha)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} - \ln\sigma\right\}.$$

La estimación $\theta^* = (\theta_1^*, \theta_2^*)$, donde $\theta_1^* = \bar{x}$, $\theta_2^* = S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - \bar{x}^2\right)$ es eficiente, puesto que pertenece a K_0, y la estadística $\left(\sum x_i, \sum x_i^2\right)$, como hemos visto en el § 15, es la estadística suficiente completa (véase el teorema 14.4).

La estimación de verosimilitud máxima $\left(\bar{x}, \frac{1}{n}\sum(x_i - \bar{x})^2\right)$ se distingue de θ^* sólo por el factor $\frac{n-1}{1}$ de la segunda coordenada, debido a lo cual la misma permanece desplazada. Para la estimación elegida θ^*, la representación exponencial especial (15) de la función $f_\theta(X)$ no se realizará, ya que

$$f_\theta(X) = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\alpha}{\sigma^2}\sum x_i - \frac{n\alpha^2}{2\sigma^2} - n\ln\sigma\right\} =$$

$$= (2\pi)^{-n/2}\exp\left\{\frac{\alpha n}{\sigma^2}\theta_1^* - \frac{n-1}{2\sigma^2}\theta_2^* - \frac{n}{2\sigma^2}(\theta_1^*)^2 - \frac{n\alpha^2}{2\sigma^2} - n\ln\sigma\right\}.$$

Esto significa que en la desigualdad multidimensional de Rao — Cramer no será alcanzada la frontera inferior.

El elipsoide de dispersión mínimo, definido (según el teorema 1A) por la matriz $I(\theta)$ (o $I^{-1}(\theta)$), se alcanzará sólo asintóticamente cuando $n \to \infty$, así que la estimación θ^*, sin ser R-eficiente, será la estimación a.R-e. Cerciorémonos de ello directamente.

Calculemos al principio la matriz $I(\theta)$. Tenemos

$$l_1'(x, \theta) = \frac{(x-\alpha)}{\sigma^2}, \quad l_2'(x, \theta) = \frac{(x-\alpha)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$$

(recordemos que l_2' no es derivada respecto a σ sino respecto a σ^2, comparen esto con el ejemplo 3). Por eso

$$I_{11}(\theta) = \mathbf{M}_\theta\frac{(x_1-\alpha)^2}{\sigma^4} = \frac{1}{\sigma^2},$$

$$I_{12}(\theta) = I_{21}(\theta) = \mathbf{M}_\theta\left[\frac{(x_1-\alpha)^3}{2\sigma^6} - \frac{x_1-\alpha}{2\sigma^4}\right] = 0,$$

$$I_{22}(\theta) = \frac{1}{4\sigma^8}\mathbf{M}_\theta[(x_1-\alpha)^2 - \sigma^2]^2 = \frac{1}{2\sigma^4}.$$

De aquí hallamos

$$(nI(\theta))^{-1} = \begin{bmatrix}\sigma^2/n & 0\\ 0 & 2\sigma^4/n\end{bmatrix}. \tag{22}$$

Calculemos ahora, para comparar, la matriz de segundos momentos centrales de la estimación θ^*.

Tenemos

$$\mathbf{M}_\theta(\theta_1^* - \theta_1)^2 = \mathbf{M}_\theta(\bar{x} - \alpha)^2 = \frac{\sigma^2}{n},$$

$$\mathbf{M}_\theta(\theta_2^* - \theta_2)^2 = \mathbf{M}_\theta(S_0^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1},$$

$$\mathbf{M}_\theta(\theta_1^* - \theta_1)(\theta_2^* - \theta_2) = 0.$$

Las dos últimas ecuaciones se calculan directamente. Examinemos, por ejemplo, la segunda de ellas. Es suficiente convencernos de que

$$\mathbf{M}_\theta(\bar{x} - \alpha)S_0^2 = 0. \tag{23}$$

Pero

$$S_0^2 = \frac{1}{n-1}\left[\sum(x_i - \alpha)^2 - (\bar{x} - \alpha)^2\right],$$

$$(\bar{x} - \alpha)S_0^2 = \frac{1}{n(n-1)}\left[\sum(x_i - \alpha)\right]\left[\sum(x_i - \alpha)^2\right] - \frac{1}{n(n-1)}(\bar{x} - \alpha)^3.$$

En vista de que

$$\mathbf{M}_\theta(\bar{x} - \alpha)^3 = \mathbf{M}_\theta(x_i - \alpha)^3 = \mathbf{M}_\theta(x_i - \alpha)(x_i - \alpha)^2 = 0,$$

(23) queda demostrada.

Ahora bien, la matriz de segundos momentos $\theta^* - \theta$ es igual a

$$\begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n-1) \end{bmatrix}.$$

Por supuesto que la diferencia entre esta matriz y la matriz $(nI(\theta))^{-1}$ puede ser considerable sólo para pequeños valores de n.

4. Algunas deducciones. Concluyendo este párrafo, hagamos cierto resumen de las investigaciones realizadas en los seis últimos párrafos. Su finalidad principal consistía en buscar los métodos de construir las estimaciones óptimas (en uno u otro sentido) y fijar las fronteras inferiores para sus desviaciones estándar. Como resultado se pueden indicar las siguientes cuatro tendencias principales de búsqueda de las mejores estimaciones.

1. Construcción de las estimaciones bayesianas (si hay una información a priori sobre θ) y minimax.

2. Determinación de las estadísticas suficientes completas (o mínimas) S. Entonces la estimación $\theta_S^* = \mathbf{M}_\theta(\theta^*/S)$ será eficiente en la clase $\mathbf{K}_b$, a la cual pertenece θ^*.

3. Utilización de las e.v.m. en los casos en que se cumple el criterio (3) del teorema 1 (o el criterio (15) del teorema 1A). En este caso también obtendremos las estimaciones eficientes (e incluso R-eficientes) en las clases con un desplazamiento registrado.

4. Enfoque cuantitativo basado en la comparación de la desviación estándar $\mathbf{M}_\theta(\theta^* - \theta)^2$ de la estimación θ^*, que queremos utilizarla, con la frontera inferior R definida por la desigualdad de Rao — Cramer. Si la relación $\mathbf{M}_\theta(\theta^* - \theta)^2/R$ es próxima a cero, la estimación θ^* puede ser recomendada para el uso. Siguiendo esta tendencia, obtendremos ulteriormente resultados muy generales relacionados con la construcción de las estimacio-

nes asintóticamente eficientes, asintóticamente bayesianas y asintóticamente minimax.

Hagamos también la siguiente observación. En todas las tendencias señaladas más arriba, desempeña un papel muy importante la forma en que la distribución de la muestra $\mathbf{P}_\theta$ depende del parámetro θ que se estima. Sin embargo, en la práctica a menudo surgen problemas de no estimación del propio θ sino de cierta función $\varphi(\theta)$ de éste. Además es facil notar (véase el ejemplo con el esquema de Bernoulli en (8.4) y (8.5)) que la estimación $\varphi^* = \varphi(\theta^*)$ no siempre, ni mucho menos, poseerá las propiedades que poseía la estimación θ^* (no estar desplazada, ser eficaz, etc., sólo se conservarán las propiedades de eficacia asintótica si φ es una función suave). Desde este punto de vista es natural que al principio se examine el problema de estimación de las *funciones* $\varphi(\theta)$ del parámetro inicial θ. Pero hemos renunciado a tal enfoque, ya que, manteniendo esta tendencia, muchos resultados básicos, obtenidos por nosotros, se complicarían considerablemente. Por otro lado, si φ realiza una aplicación biunívoca, el problema de estimación de $\varphi(\theta)$ se reducirá al problema examinado por nosotros mediante la «reparametrización», o sea, la introducción de un nuevo parámetro $\gamma = \varphi(\theta)$, al que le corresponderá la familia de distribuciones $\mathbf{G}_\gamma = \mathbf{P}_\varphi - 1_{(\gamma)}$.

§ 17*. Propiedades de la información de Fisher

Ya hemos visto, y nos convenceremos en adelante, que la información de Fisher desempeña un papel muy importante en la estadística matemática. Por eso aclaremos algunas propiedades útiles de la misma.

1. Caso unidimensional. La información de Fisher,

$$I(\theta) = \int \frac{(f'_\theta(x))^2}{f_\theta(x)}\,\mu(dx) = \mathbf{M}_\theta(l'(\mathrm{x}_1, \theta))^2,$$

apareció en las investigaciones del párrafo precedente. La magnitud

$$I^X(\theta) = \mathbf{M}_\theta[L'(X, \theta)]^2$$

suele considerarse como la *medida de la cantidad de información contenida en la muestra X respecto al parámetro* θ. En el teorema 16.1 hemos demostrado la *aditividad* de la información: $I^n(\theta) = nI(\theta)$, o sea, que $I^n(\theta)$ es igual a la suma de informaciones $I^{\mathrm{x}_i}(\theta) = \mathbf{M}_\theta[l'(\mathrm{x}_i, \theta)]^2 = I(\theta)$ contenidas en las observaciones independientes $\mathrm{x}_1, \ldots, \mathrm{x}_n$.

Demostremos una propiedad más de la información de Fisher. Sea $S = S(X)$ cierta estadística con valores en R^l, y sea $g_\theta(s)$ la densidad de su distribución inducida por la distribución $\mathbf{P}_\theta$ en $(\mathscr{X}^n, \mathfrak{B}_{\mathscr{X}^n})$ respecto a cierta medida λ en $(R^l, \mathfrak{B}^l)$. De acuerdo con las designaciones anteriores,

llamaremos la magnitud

$$I^S(\theta) = \mathbf{M}_\theta[(\log g_\theta(S))']^2$$

información contenida en la estadística S respecto al parámetro θ.

Notemos que el valor de $I^S(\theta)$ no depende de la elección de la medida λ. En efecto, si $\tilde{\lambda}$ es cualquier otra medida y $\nu = \lambda + \tilde{\lambda}$. Entonces λ y $\tilde{\lambda}$ serán absolutamente continuas respecto a ν, y la densidad $g^\nu_\theta(s)$ de la distribución de S respecto a la medida ν será igual a

$$g^\nu_\theta(s) = g_\theta(s)\,\frac{d\lambda}{d\nu} = \tilde{g}_\theta(s)\frac{d\tilde{\lambda}}{d\nu},$$

donde $\tilde{g}_\theta$ es la densidad respecto a $\tilde{\lambda}$. Como $\frac{d\lambda}{d\nu}$ y $\frac{d\tilde{\lambda}}{d\nu}$ no dependen de θ, las derivadas de los logaritmos de todas las tres expresiones coincidirán.

Teorema 1. *Supongamos que las densidades $f_\theta(x)$ y $g_\theta(s)$ satisfacen las condiciones (R). Entonces*

$$I^S(\theta) \leqslant I^X(\theta). \tag{1}$$

Aquí la igualdad se alcanza si y sólo si S es una estadística suficiente.

Demostración. Para cualquier $B \in \mathfrak{B}^l$ designemos por $S^{-1}(B) \in \mathfrak{B}^n_{\mathscr{X}}$ el conjunto $x \in \mathscr{X}^n$ para el cual $S(x) \in B$. Entonces según la definición de la e.m.c.,

$$\int_{S^{-1}(B)} L'(x,\,\theta)\mathbf{P}_\theta(dx) = \mathbf{M}_\theta[L'(X,\,\theta);\; X \in S^{-1}(B)] =$$

$$= \mathbf{M}_\theta[\mathbf{M}_\theta(L'(X,\,\theta)/S);\; S \in B]. \tag{2}$$

Por otro lado,

$$\int_{S^{-1}(B)} L'(x,\,\theta)\mathbf{P}_\theta(dx) = \frac{\partial}{\partial\theta}\int_{S^{-1}(B)} f_\theta(x)\mu^n(dx) = \frac{\partial}{\partial\theta}\int_B g_\theta(s) \times$$

$$\times\,\lambda(ds) = \int_B \frac{\partial}{\partial\theta}\, g_\theta(s)\lambda(ds) = \mathbf{M}_\theta[(\log g_\theta(S))';\; S \in B]. \tag{3}$$

Comparando (2) y (3), vemos que c.d. $[\mathbf{P}_\theta]$

$$\mathbf{M}_\theta(L'(X,\,\theta)/S) = (\log g_\theta(S))'. \tag{4}$$

Luego tenemos

$$0 \leqslant \mathbf{M}_\theta[L'(X,\,\theta) - (\log g_\theta(S))']^2 =$$

$$= I^X(\theta) + I^S(\theta) - 2\mathbf{M}_\theta L'(X,\,\theta)(\log g_\theta(S))',$$

donde, en virtud de (4),

$$\mathbf{M}_\theta L'(X, \theta)(\log g_\theta(S))' =$$
$$= \mathbf{M}_\theta[(\log g_\theta(S))'\mathbf{M}_\theta(L'(X, \theta)/S] = \mathbf{M}_\theta[(\log g_\theta(S))']^2 = I^S(\theta).$$

Esto demuestra la desigualdad (1).

Sea ahora S una estadística suficiente para θ. Entonces

$$f_\theta(X) = \psi(S, \theta)h(X). \tag{5}$$

Tomemos en calidad de λ la medida

$$\lambda(B) = \int_{S^{-1}(B)} h(x)\mu^n(dx).$$

Entonces, como se muestra en el lema 15.1, la distribución de S será absolutamente continua respecto a λ y tendrá una densidad $g_\theta(s)$ igual a $g_\theta(s) = \psi(s, \theta)$. De aquí, en virtud de (5), obtenemos

$$I^X(\theta) = \mathbf{M}[L'(X, \theta)]^2 = \mathbf{M}_\theta[(\log \psi(S, \theta))']^2 = Y^S(\theta).$$

Mostremos ahora que de todas las igualdades $Y^X(\theta) = Y^S(\theta)$ para todos θ se deduce que S es estadística suficiente. Efectivamente, $Y^X(\theta)$ es la dispersión de $L'(X, \theta)$, así que

$$I^X(\theta) = \mathbf{M}_\theta[L'(X, \theta) - \mathbf{M}_\theta(L'(X, \theta)/S)]^2 + \mathbf{M}_\theta[\mathbf{M}_\theta(L'(X, \theta)/S]^2. \tag{6}$$

Pero, en virtud de (4), el último sumando es igual a

$$\mathbf{M}_\theta[(\log g_\theta(S))']^2 = I^S(\theta).$$

Como $I^X(\theta) = I^S(\theta)$, entonces en (6) c.d. $[\mathbf{P}_\theta]$ para todos θ,

$$L'(X, \theta) - \mathbf{M}_\theta(L'(X, \theta)/S) = 0.$$

Por lo tanto, $L'(X, \theta)$ es medible respecto a $\sigma(S)$ y, por consiguiente, existe una función medible $\varphi(S, \theta)$ tal que

$$L'(X, \theta) = \varphi(S, \theta), \; L(X, \theta) = \Phi(S, \theta) + h_1(X),$$
$$f_\theta(X) = \exp[\Phi(S, \theta) + h_1(X)]. \; \triangleleft$$

Ya hemos señalado que las estadísticas suficientes son el tipo único de estadísticas que reducen los datos muestrales sin perder la información acerca del parámetro θ. El teorema 1 confiere a esta afirmación el sentido exacto con arreglo a la información de Fisher.

Ejemplo 1. Sea $X \in \mathbf{B}_p$. Aquí

$$f_p(x) = p^x(1 - p)^{1-x},$$

donde x es igual a 0 ó a 1, y $f_p(x)$ es la densidad respecto a la medida

de cálculo. Por eso

$$l(x,\ p) = x \ln p + (1 - x)\ln(1 - p),$$

$$l'(x,\ p) = \frac{x}{p} - \frac{1 - x}{1 - p},$$

$$I(p) = \mathbf{M}_p[l'(x_1,\ p)]^2 = p\left(\frac{1}{p}\right)^2 + (1 - p)\left(\frac{1}{1 - p}\right)^2 = \frac{1}{p(1 - p)}.$$

Ahora bien, la información de una observación en el esquema de Bernoulli es igual a $(p(1 - p))^{-1}$ y alcanza su valor mínimo cuando $p = 1/2$.

La información de toda la muestra constituye $n/(p(1 - p))$. Designemos ahora por ν el número de «casos favorables» en la muestra X (número de casos unitarios) y hallemos la información de esta observación. Las densidades (otra vez respecto a la medida de cálculo) para ν serán iguales a

$$g_p(x) = C_n^x p^x (1 - p)^{n - x},\ x = 0,\ \ldots,\ n,$$

así que $\log g_p(x) = x \log p + (n - x)\log(1 - p) + \log C_n^x$,

$$I^{\nu}(p) = \mathbf{M}_p[(\log g_p(\nu))']^2 =$$

$$= \sum_{x=0}^{n} C_n^x p^x (1 - p)^{n - x}\left(\frac{x}{p} - \frac{n - x}{1 - p}\right)^2 = \sum_{x=0}^{n} C_n^x p^x (1 - p)^{n - x} \times$$

$$\times \frac{(x - np)^2}{(p(1 - p))^2} = \frac{1}{(p(1 - p))^2}\,\mathbf{D}\nu = \frac{n}{p(1 - p)}.$$

Esta igualdad concuerda por completo con el teorema 1.

Le proponemos al lector que halle, en forma de ejercicios, las informaciones de observaciones para las muestras de las distribuciones que dependen del parámetro unidimensional y que han sido dadas en el § 2.

2. Caso multidimensional. Sea ahora $\theta \in R^k$, $k > 1$. En este caso se trata de la matriz de información de Fisher de la observación x_1:

$$I(\theta) = \|I_{ij}(\theta)\|,\ I_{ij}(\theta) = \mathbf{M}_\theta \frac{\partial}{\partial \theta_i} l(x_1,\ \theta) \frac{\partial}{\partial \theta_j} l(x_1,\ \theta),$$

donde se supone, claro está, que la función $f_\theta(x)$ es derivable.

Si ponemos

$$\varphi(x,\ \theta) = (\varphi_1(x,\ \theta),\ \ldots,\ \varphi_k(x,\ \theta)) =$$

$$= 2(\sqrt{f_\theta(x)})' = \frac{1}{\sqrt{f_\theta(x)}}\left(\frac{\partial f_\theta(x)}{\partial \theta_1},\ \ldots,\ \frac{\partial f_\theta(x)}{\partial \theta_k}\right),$$

entonces la matriz $I(\theta)$ también puede ser escrita en la forma

$$I(\theta) = \int_{\mathscr{X}} \varphi^T(x,\ \theta)\varphi(x,\ \theta)\mu(dx).$$

Ya hemos establecido, en el § 16, que al igual que en el caso unidimensional, la información de Fisher es aditiva, o sea, la matriz de información de Fisher de la muestra X es igual a la suma de las matrices de información de distintas observaciones. Si designamos

$$I^X(\theta) = \|I_{ij}^X(\theta)\|,\ I_{ij}^X(\theta) = \mathbf{M}_\theta \frac{\partial}{\partial\theta_i} L(X, \theta) \frac{\partial}{\partial\theta_j} L(X, \theta),$$

entonces $I^X(\theta) = nI(\theta)$.

El teorema 1 también es completamente válido. Sea $g_\theta(s)$ la densidad de cierta estadística $S = S(X)$ con valores en R^l respecto a cierta medida λ. Designemos

$$I^S(\theta) = \|I_{ij}^S(\theta)\|,\ I_{ij}^S(\theta) = \mathbf{M}_\theta \frac{\partial}{\partial\theta_i} \log g_\theta(S) \frac{\partial}{\partial\theta_j} \log g_\theta(S).$$

Hemos obtenido la matriz de información de la observación S.

Teorema 1A. *Si las densidades $f_\theta(x)$ y $g_\theta(s)$ satisfacen las condiciones (R) del* § 16, *entonces*

$$I^S(\theta) \leqslant I^X(\theta), \tag{7}$$

o sea, la matriz $I^X(\theta) - I^S(\theta)$ es definida no negativamente. La igualdad en (7) *tiene lugar si y sólo si S es una estadística suficiente.*

La **demostración** de este teorema es completamente análoga a la del teorema 1 y, para abreviar, la omitimos. La misma se puede hallar, por ejemplo, en [95] y [48].

Ejemplo 2. En el § 16 ya hemos calculado la matriz de información para una distribución normal. Calculémosla ahora para una familia biparamétrica de distribuciones

$$f_\theta(x) = \frac{1}{\sigma} f\left(\frac{x - \alpha}{\sigma}\right),$$

donde $\theta = (\alpha, \sigma)$, f es una función derivable dada, para la cual existen las integrales

$$I_i = \int x^i \frac{(f'(x))^2}{f(x)}\, dx = \mathbf{M}_{(0,1)} x_1^i (l'(x_1))^2,\ i = 0, 1, 2.$$

Aquí $l(x) = \log f(x)$; la tilde $'$ significa la derivación ordinaria, y α y σ son los parámetros de desplazamiento y escala de una distribución de densidad $f(x)$. Ahora bien, conocemos el tipo de la distribución, pero sólo con una exactitud de hasta la transformación lineal del argumento. Los parámetros α y σ de la distribución normal Φ_{α,σ^2} son, evidentemente, los parámetros de desplazamiento y escala. Al ser registrado λ, el parámetro λ de la

distribución Γ es un parámetro de escala, al igual que el parámetro θ en la distribución $\mathbf{U}_{0,\theta}$.

Tenemos

$$l(x,\ \theta) = \log f_\theta(x) = -\log\sigma + l\left(\frac{x-\alpha}{\sigma}\right),$$

$$\frac{\partial l(x,\ \theta)}{\partial\alpha} = -\frac{1}{\sigma}\, l'\left(\frac{x-\alpha}{\sigma}\right),$$

$$\frac{\partial l(x,\ \theta)}{\partial\sigma} = -\frac{1}{\sigma} - \frac{(x-\alpha)}{\sigma^2}\, l'\left(\frac{x-\alpha}{\sigma}\right).$$

De aquí hallamos

$$I_{11}(\theta) = \frac{1}{\sigma^2}\,\mathbf{M}_\theta\left[l'\left(\frac{x_1-\alpha}{\sigma}\right)\right]^2 = \frac{1}{\sigma^2}\int \frac{\left[f'\left(\frac{x-\alpha}{\sigma}\right)\right]^2}{f\left(\frac{x-\alpha}{\sigma}\right)}\,dx = \frac{1}{\sigma^2}\,I_0,$$

$$I_{12}(\theta) = \frac{1}{\sigma^2}\,\mathbf{M}_\theta l'\left(\frac{x_1-\alpha}{\sigma}\right)\left[1 + \frac{x_1-\alpha}{\sigma}\, l'\left(\frac{x_1-\alpha}{\sigma}\right)\right] = \frac{1}{\sigma^2}\,I_1,$$

$$I_{22}(\theta) = \frac{1}{\sigma^2}\,\mathbf{M}_\theta\left[1 + \frac{x_1-\alpha}{\sigma}\, l'\left(\frac{x_1-\alpha}{\sigma}\right)\right]^2 = \frac{1}{\sigma^2}\,[I_2 - 1],$$

puesto que $2\int \frac{x-\alpha}{\sigma}\, f'\left(\frac{x-\alpha}{\sigma}\right)\frac{dx}{\sigma} = -2\int f(x)dx = -2$. Por lo tanto,

$$I(\theta) = \frac{1}{\sigma^2}\begin{Vmatrix} I_0 & I_1 \\ I_1 & I_2 - 1 \end{Vmatrix}.$$

Si f es una función simétrica, es evidente que $I_1 = 0$.

La degeneración de la matriz $I(\theta)$ significa que su determinante se reduce a cero o, que es lo mismo,

$$[\mathbf{M}_{(0,1)}l'(x_1)(1 + x_1 l'(x_1))]^2 = \mathbf{M}_{(0,1)}(l'(x_1))^2\mathbf{M}_{(0,1)}(1 + x_1 l'(x_1))^2.$$

Esto es posible únicamente en el caso cuando $1 + xl'(x) = cl'(x)$ para cualquier c, o cuando $l'(x) = 0$. De la primera igualdad se deduce que

$$l(x) = -\ln(x-c) + c_1,\ f(x) = \frac{e^{c_1}}{x-c}.$$

Está claro que tal función $f(x)$ no puede ser la densidad de la distribución. Análogamente se examina la posibilidad de que $l'(x) = 0$. Por lo tanto, $I(\theta)$ está definida positivamente.

En particular, para la familia normal $\{\Phi_{\alpha,\sigma^2}\}$, cuando $\theta = (\alpha, \sigma)$,

$$I(\theta) = \frac{1}{\sigma^2} \begin{Vmatrix} 1 & 0 \\ 0 & 2 \end{Vmatrix},$$

puesto que en este caso $l(x) = -x^2/2 - \ln\sqrt{2\pi}$, $l'(x) = -x$, $I_0 = \mathbf{M}_{(0,1)}x_1^2 = 1$, $I_1 = \mathbf{M}_{(0,1)}x_1^3 = 0$, $I_2 = \mathbf{M}_{(0,1)}x_1^4 = 3$. Podríamos haber obtenido este mismo resultado con ayuda del ejemplo 16.4, si hubiéramos utilizado los datos del apartado 3 donde hemos mostrado el comportamiento de la matriz de información al sustituir el parámetro (en el ejemplo 16.4 $\theta = (\alpha, \sigma^2)$, pero no (α, σ)). Le proponemos al lector que se cerciore de que, en concordancia con el teorema 1A, la estadística $\left(\bar{x}, \sum x_i^2\right)$ tiene la matriz de información

$$I^S(\theta) = \frac{1}{\sigma^2} \begin{Vmatrix} 1 & 0 \\ 0 & 2 \end{Vmatrix} = nI(\theta).$$

3. Matriz de Fisher y sustitución del parámetro. Examinemos la cuestión de cómo se comporta la matriz de información al sustituir el parámetro. Pongamos $\theta = v(\beta)$, $\beta \in R^k$, donde v es una función vectorial derivable, y examinemos la familia paramétrica $\mathbf{P}_\beta^{(1)} = \mathbf{P}_{v(\beta)}$. Con el fin de hallar la matriz de información $J(\beta)$ para esta familia, debemos hallar las derivadas

$$\frac{\partial}{\partial \beta_j} l(x_1, v(\beta)) = \sum_{i=1}^{k} \frac{\partial}{\partial \theta_i} l(x_1, v(\beta)) \frac{\partial v_i(\beta)}{\partial \beta_j}. \tag{8}$$

Si designamos $V = \left\| \frac{\partial v_i(\beta)}{\partial \beta_j} \right\|$, $i, j = 1, \ldots, k$, obtenemos que el vector de las derivables en (8) $l_\theta'(x_1, v(\beta))$ es representable en la forma $l_\theta'(x_1, v(\beta))V$, así que

$$J(\beta) = \mathbf{M}_\beta(l_\theta'(x_1, v(\beta))V)^T(l_\theta'(x_1, v(\beta))V) = V^T I(v(\beta))V.$$

En particular, si $\theta = \beta C$, $C = \|c_{ij}\|$, $i, j, = 1, \ldots, k$, entonces $V = C^T$ y

$$J(\beta) = CI(\theta)C^T. \tag{9}$$

Obsérvese que si examinamos, en el espacio paramétrico, el elipsoide

$$(\theta - \theta_1)I(\theta)(\theta - \theta_1)^T < c, \tag{10}$$

la escritura (10) de este conjunto es invariante con respecto a la transformación invertible lineal C sobre el parámetro θ. Así pues, si ponemos $\theta = \beta C$, el conjunto (10) en nuevas variables tendrá la forma

$$(\beta - \beta_1)J(\beta)(\beta - \beta_1)^T < c,$$

donde $\beta_1 = \theta_1 C^{-1}$. Esto se obtiene inmediatamente si se sustituye $\theta = \beta C$ en (10) y si utilizamos (9).

§ 18*. Estimaciones del parámetro de desplazamiento y escala. Estimaciones equivariantes eficientes

En los §§ 12—16 hemos visto y nos convenceremos posteriormente hasta qué punto es útil el concepto de estadística suficiente en general y al construir las estimaciones eficientes en particular. El círculo de ideas relacionadas con la utilización de las estadísticas suficientes podría llamarse *principio de suficiencia.* Al construir las estimaciones eficientes hemos combinado el principio de suficiencia con otro principio llamado *principio de no desplazamiento.* Este último consiste en separar las clases de estimaciones con desplazamento registrado y, en particular, con desplazamiento nulo. Sin registrar el desplazamiento sería imposible separar las estimaciones eficientes.

En este párrafo, así como en el párrafo siguiente y en el capítulo 3, examinaremos el tercer principio importante de la estadística matemática, o sea, el *principio de invariación.*

La introducción de todos los principios mencionados tiene el mismo sentido: ellos permiten, de un modo natural, reducir la clase de las estimaciones sujetas a estudio, de manera que en las reducciones obtenidas resulte posible la determinación de las estimaciones eficientes.

1. Estimaciones del parámetro de desplazamiento y escala. Se llama problema de estimación del parámetro de desplazamiento el problema de estimación del parámetro α en la familia de distribuciones $\{\mathbf{P}_\alpha\}$ que poseen la propiedad

$$\mathbf{P}_\alpha(A) = \mathbf{P}(A - \alpha).$$

Aquí $\mathbf{P}$ es cierta distribución registrada; $A - \alpha = \{x : x + \alpha \in A\}$ y se supone que el conjunto paramétrico Θ tiene la misma naturaleza que $\mathscr{X}$. En el caso en que $\mathscr{X} = R^m$ se puede, por supuesto, examinar también los desplazamientos de θ de "menor dimensión", por ejemplo, escalares, pero entonces es necesario registrar la dirección (vector $e \in \mathscr{X}$) de desplazamiento y estudiar $\mathbf{P}_\alpha(A) = \mathbf{P}(A + \alpha e)$. Para abreviar, examinaremos tan sólo la primera posibilidad y consideraremos que $\Theta = \mathscr{X} = R^m$.

Señalemos que la distribución $\mathbf{P}_\alpha$ de $x_i + c (c \in R^m)$ coincide con la distribución $\mathbf{P}_{\alpha+c}$ de la magnitud x_i, o sea, el desplazamiento de todas las observaciones en c conduce a la muestra de la distribución $\mathbf{P}_{\alpha+c}$. Por eso es natural que se investiguen únicamente las estimaciones $\alpha^* = \alpha^*(X)$ del parámetro α que poseen la propiedad

$$\alpha^*(X + c) = \alpha^*(X) + c. \tag{1}$$

De aquí en adelante $X + c$ significará el vector con coordenadas $(x_1 + c, \ldots, x_n + c)$. La violación de esta igualdad significaría que la estimación

α^* depende del origen, o sea, de la elección del origen de coordenadas en el espacio $\mathscr{X} = R^m$.

El enfoque análogo aparece al estimar el parámetro de escala cuando se aprecia el parámetro σ en la familia $\{\mathbf{P}_\sigma\}$ que tiene la propiedad $\mathbf{P}_\sigma(A) = (A/\sigma)$, $\sigma \in (0, \infty)$. Aquí suponemos que σ es escalar, aunque se puede examinar también un caso matricial. En este caso la distribución $\mathbf{P}_\sigma$ de los valores $x_i c$ coincide con la distribución $\mathbf{P}_{\sigma c}$ de las magnitudes x_i, o sea, la multiplicación de las observaciones por c conduce a la muestra de $\mathbf{P}_{\sigma c}$. Por consiguiente, en este caso es natural limitarse al examen de las estimaciones que poseen la propiedad

$$\sigma^*(Xc) = c\sigma^*(X), \tag{2}$$

donde $Xc = (x_1 c, \ldots, x_n c)$, puesto que al variar c veces la escala de observaciones esa misma cantidad de veces también varía el parámetro.

El lector, por su propia iniciativa, puede obtener fácilmente las afirmaciones siguientes.

Si la familia $\mathbf{P}_\theta$ satisface la condición (A_μ), entonces θ será de parámetro de desplazamiento (de escala) si y sólo si

$$f_\theta(x) = f(x - \theta), \quad \left(f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)\right).$$

Si $\mathscr{X} \supset R = \Theta$, $X \in \mathbf{P}_\alpha$ y α es el parámetro de desplazamiento, entonces $Y = e^X = (e^{x_1}, \ldots, e^{x_n}) \in \mathbf{Q}_\sigma$, donde, para las distribuciones $\mathbf{Q}_\sigma$, $\sigma = e^\alpha$ es el parámetro de escala. Esto se deduce directamente del hecho de que la densidad $y_1 = e^{x_1}$ es igual a (véase [11], p.)

$$\frac{1}{y} f(\ln - \alpha) = \frac{1}{\sigma}\left[\frac{\sigma}{y} f\left(\ln \frac{y}{\sigma}\right)\right].$$

Al contrario, si $\mathscr{X} = (0, \infty) = \Theta$, $X \in \mathbf{P}_\sigma$ y σ es el parámetro de escala, entonces $Y = \ln X = (\ln x_1, \ldots, \ln x_n) \in \mathbf{Q}_\alpha$, donde $\alpha = \ln \sigma$ es el parámetro de desplazamiento de las distribuciones $\mathbf{Q}_\alpha$.

Se puede examinar también el problema de estimación simultánea de los parámetros desconocidos α y σ en el caso en que $\mathbf{P}_{\alpha,\sigma}(A) = \mathbf{P}\left(\frac{A - \alpha}{\sigma}\right)$. En estas condiciones es natural que en calidad de estimación de σ se examinen las funciones que poseen la propiedad

$$\alpha^*(X + c) = \alpha^*(X), \quad \sigma^*(Xc) = c\sigma^*(X). \tag{3}$$

Las estimaciones que en los ejemplos examinados satisfacen las condiciones (1), (2) y (3) se llaman *equivariantes* (véase la definición general en el § 19). La causa de introducción de tales estimaciones consiste en la con-

tracción de todas las estimaciones sometidas a estudio, lo cual simplifica el problema de búsqueda de las estimaciones óptimas. Así en el § 8 hemos establecido que es imposible hallar uniformemente (o sea, para todos los θ) las mejores estimaciones en la clase de todas las estimaciones. Resulta que en la clase de estimaciones equivariantes tales estimaciones uniformemente mejores ya existen y en varios casos pueden ser halladas en forma explícita. Vamos a ilustrar este hecho citando, a título de ejemplo, las estimaciones de desplazamiento y escala.

2. Estimación eficiente del parámetro de desplazamiento en la clase de estimaciones equivariantes. Aquí consideraremos que se cumple la condición (A_μ) y, por lo tanto, $f_\alpha(x) = f(x - \alpha)$ y que μ es la medida de Lebesgue.

Designemos por S_0 la estadística

$$S_0 = S_0(X) = (x_2 - x_1, \ \ldots, \ x_n - x_1)$$

que es, evidentemente, invariante respecto al desplazamiento: $S_0(X + c) = S_0(X)$. Designemos por K_E la clase de todas las estimaciones equivariantes α^*, o sea, las estimaciones que satisfacen (1), y designemos por $|\alpha|^2$ el cuadrado de la norma euclídea $\alpha \in R^m$.

Teorema 1. *Sea $\alpha^* = \alpha^*(X)$ cualquier estimación equivariante con valor finito $\mathbf{M}_0\alpha^*$. Entonces, la estimación*

$$\alpha_0^* = \alpha^* - \mathbf{M}_0(\alpha^*/S_0) \tag{4}$$

no depende de la elección de α^* y es la única estimación eficiente en la clase K_E, o sea, $\mathbf{M}_\alpha|\alpha_0^* - \alpha|^2 = \min_{\alpha^* \in K_E} \mathbf{M}_\alpha|\alpha^* - \alpha|^2$ para todos los α y $\mathbf{M}_\alpha|\alpha^* - \alpha|^2 = \mathbf{M}_\alpha|\alpha_0^* - \alpha|^2$ si sólo $\mathbf{M}_0(\alpha^*/S_0) = 0$ c.d. La estimación α_0^* puede ser representada en la forma

$$\alpha_0^* = \frac{\int u f_u(X)du}{\int f_u(X)du} = \frac{\int u f(X - u)du}{\int f(X - u)du}. \tag{5}$$

La estimación α_0^* se denomina *estimación de Pitman*. De (4) es fácil deducir que ésta es equivariante y no está desplazada. La equivariación se deduce de la equivariación de α^* y de la invariación respecto al desplazamiento de la función $V(S_0) = \mathbf{M}_0(\alpha^*/S_0)$ que depende tan sólo de S_0. El no desplazamiento se deduce de las igualdades

$$\mathbf{M}_\alpha\alpha_0^* = \alpha + \mathbf{M}_\alpha\alpha^*(X - \alpha) - \mathbf{M}_\alpha V(S_0), \tag{6}$$

donde $\mathbf{M}_\alpha V(S_0) = \mathbf{M}_0 V(S_0)$, $\mathbf{M}_\alpha\alpha^*(X - \alpha) = \mathbf{M}_0\alpha^*(X)$. La última relación se deduce del hecho de que $X - \alpha \in \mathbf{P}_0$ si $X \in \mathbf{P}_\alpha$. Por eso la suma de los dos últimos sumandos en (6) constituye

$$\mathbf{M}_0\alpha^* - \mathbf{M}_0[\mathbf{M}_0(\alpha^*/S_0)] = 0; \ \mathbf{M}_\alpha\alpha_0^* = \alpha.$$

Antes de demostrar el teorema expondremos la siguiente afirmación auxiliar.

Lema 1. *Sea* $X \in \mathbf{P}_0$. *Para cualquier estadística* $S = S(X)$ *con esperanza matemática finita* $\mathbf{M}_0|S| < \infty$, *la e.m.c. de* S *respecto a* S_0 *es igual a*

$$\mathbf{M}_0(S/S_0) = S_1(X) \equiv \frac{\int S(X-u)f_u(X)du}{\int f_u(X)du}. \tag{7}$$

Demostración. Todas las funciones bajo los signos integrales en (7) son las funciones de $X - u$. Por consiguiente, después de sustituir $x_1 - u = v$, las mismas serán las funciones de $(v, x_2 - x_1 + v, \ldots, x_n - x_1 + v)$. Esto quiere decir que el segundo miembro de (7) depende únicamente de S_0. En virtud de las propiedades de la e.m.c., para demostrar el lema es suficiente convencerse que para cualquier $A \in \sigma(S_0)$

$$\mathbf{M}_0(S_1;\ A) = \mathbf{M}_0(S;\ A). \tag{8}$$

Sea $Z = Z(S_0)$ cualquier estadística $\sigma(S_0)$-medible limitada. Entonces

$$\mathbf{M}_0 Z S_1 = \int_{\mathscr{X}^n} \frac{Z(S_0)\int_\Theta S(x-u)f_u(x)du}{\int_\Theta f_u(x)du} f(x)\,dx =$$

$$= \int_\Theta \int_{\mathscr{X}^n} \frac{Z(S_0)S(x-u)f(x-u)f(x)}{\int_\Theta f(x-v)dv}\,dx\,du.$$

Después de sustituir $x - u \to x$, en el intervalo interior obtenemos (en este caso $S_0(x)$ se transforma en sí mismo)

$$\int_\Theta \int_{\mathscr{X}^n} \frac{Z(S_0)S(x)f(x)f(x+u)}{\int_\Theta f(x+u-v)dv}\,dx\,du = \int_{\mathscr{X}^n} Z(S_0)S(x)f(x)dx = \mathbf{M}_0 ZS.$$

Esto demuestra (8). El cambio del orden de integración, al cual hemos acudido dos veces, es justo en virtud de la integrabilidad absoluta de S y del carácter limitado de Z. ◁

Demostración del teorema 1. Antes que nada es preciso señalar que para la estimación equivariante, $\mathbf{M}_\alpha|\alpha^* - \alpha|^2$ no depende de α. En efecto,

$$\mathbf{M}_\alpha|\alpha^*(X) - \alpha|^2 = \mathbf{M}_\alpha|\alpha^*(X-\alpha)^2 = \mathbf{M}_0|\alpha^*(X)|^2.$$

Ahora bien, para determinar la estimación equivariante uniformemente óptima es necesario hallar α^*, que minimiza $\mathbf{M}_0|\alpha^*|^2$.

Sea α^* cualquier estimación equivariante α. En virtud de las propiedades de la e.m.c.,

$$\mathbf{M}_0|\alpha^*|^2 = \mathbf{M}_0|\alpha^* - \mathbf{M}_0(\alpha^*/S_0)|^2 + \mathbf{M}_0|\mathbf{M}_0(\alpha^*/S_0)|^2 \geqslant \geqslant \mathbf{M}_0|\alpha^* - \mathbf{M}_0(\alpha^*/S_0)|^2. \qquad (9)$$

Queda señalar que, en virtud del lema 1, la estimación $\alpha_0^* = \alpha^* - \mathbf{M}_0(\alpha^*/S_0)$ es igual a (5) y no depende de la elección de α^*. La igualdad en (9) es, evidentemente, posible si y sólo si $\mathbf{M}_0(\alpha^*/S_0) = 0$ c.d. ◁

De la demostración del teorema se deduce que, en la construcción de la estimación óptima equivariante, desempeña un papel especial la estadística $S_0 = (x_2 - x_1, \ldots, x_n - x_1)$, que es invariante respecto a la transformación del desplazamiento. La invariación de la estadística es, en cierto sentido, una cualidad contraria a la suficiencia, y la construcción de la estimación $\theta_0^* = \theta^* - \mathbf{M}_0(\theta^*/S_0)$ a base de la estimación arbitraria θ^*, es el enfoque del mejoramiento de la estimación θ^*, también, en cierto sentido, contrario al enfoque con el cual, para el mejoramiento de la estimación θ^* mediante la estadística suficiente S, se examina la estimación $\theta_s^* = \mathbf{M}_\theta(\theta^*/S)$. La contrariedad consiste en lo siguiente. La característica suficiente contiene toda la información sobre el parámetro θ, mientras que la estadística invariante no contiene ninguna. Con el fin de obtener las mejores estimaciones, hemos buscado las estadísticas suficientes mínimas; aquí, como veremos, necesitamos las estadísticas invariantes máximas (tal es la estadística S_0). La estimación θ_s^* es la «proyección» de θ^* sobre S, mientras que la estimación θ_0^* se obtiene sustrayendo de θ^* su «proyección» sobre S_0.

En resumidas cuentas, los resultados obtenidos por estas dos vías coinciden a menudo, como se verá de los dos ejemplos siguientes.

Ejemplo 1. Sea $\mathscr{X} = R$, $X \in \Phi_{\alpha,1}$. Entonces

$$f_\alpha(X) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i - \alpha)^2\right\} = \frac{1}{\sqrt{n}\,(2\pi)^{\frac{n-1}{2}}} \exp\left\{-\frac{1}{2}\sum(x_i - \bar{x})^2\right\} \cdot \sqrt{\frac{n}{2\pi}}\, e^{-\frac{n}{2}(\alpha - \bar{x})^2}.$$

Aquí el segundo factor, como función de α, es la función de densidad de la ley normal con parámetros $(\bar{x}, 1/n)$. Como el primer factor no depende de α, es reducido en (5), y la estimación de Pitman constituirá $\alpha^* = x$. En el caso multidimensional obtendremos este mismo resultado.

Ejemplo 2. Sea $\mathscr{X} = R$, $X \in \mathbf{U}_{\theta,1/\theta}$. Entonces

$$f_\theta(X) = \begin{cases} 1 & \text{cuando } x_{(n)} - 1 \leqslant \theta \leqslant x_{(1)}, \\ 0 & \text{en los demás casos.} \end{cases}$$

Por eso

$$\theta^* = \int_{x_{(n)}-1}^{x_{(1)}} u\,du \Big/ (x_{(1)} - x_{(n)} + 1) = \frac{1}{2}\,(x_{(1)} + x_{(n)} - 1).$$

Ahora bien, vemos que en la clase K_E de estimaciones equivariantes se pueden construir, en forma explícita, las estimaciones eficientes, además, en este caso no se necesitan ningunas condiciones de suavidad de $f_\theta(x)$, y la propia eficacia tiene un carácter exacto (no asintótico).

3. Carácter minimax de la estimación de Pitman. Ahora prestemos atención a la forma de estimación de Pitman. Hablando en términos generales, ésta es una estimación bayesiana para la distribución a priori «uniforme en todo el eje». Como tal distribución no existe, enunciemos más exactamente la referida afirmación. Sea $\mathscr{X} = R$ y $\mathbf{Q}^{(N)}$ una distribución uniforme en $[-N, N]$, o sea, una distribución cuya densidad constituye

$$q^{(N)}(t) = \begin{cases} (2N)^{-1}, & |t| \leqslant N, \\ 0 & |t| > N. \end{cases}$$

La estimación bayesiana correspondiente a $\mathbf{Q}^{(N)}$ será igual a

$$\alpha^*_{\mathbf{Q}^{(N)}} = \frac{\int u q^{(N)}(u) f_u(X)\,du}{\int q^{(N)}(u) f_u(X)\,du} = \int_{-N}^{N} u f_u(X)\,du \Big/ \int_{-N}^{N} f_u(X)\,du.$$

Es evidente que para todos X, la estimación de Pitman α^*_0 es el límite $\alpha^*_0 = \lim\limits_{N\to\infty} \alpha^*_{\mathbf{Q}}(N)$. Esta circunstancia sugiere que a la vez convergerán también los momentos de segundo orden:

$$\mathbf{M}_\alpha(\alpha^*_{Q^{(N)}} - \alpha)^2 \to \mathbf{M}_\alpha(\alpha^*_0 - \alpha)^2.$$

Resulta que en la región $|\alpha| \leqslant N - \sqrt{N}$, eso es precisamente así. Además, la convergencia será uniforme respecto a α en el referido intervalo de valores de α. (La demostración está relacionada con la estimación de $\mathbf{M}_\alpha(\alpha^*_0 - \alpha^*_{Q^{(N)}})^2$, tiene principalmente carácter técnico y por eso la omitimos).

Pero en este caso podemos utilizar el criterio del carácter minimax de las estimaciones en el teorema 11.3; *si la estimación α^* es tal que, para todos los valores de α,*

$$\mathbf{M}_\alpha(\alpha^* - \alpha)^2 \leqslant \lim_{N\to\infty} \sup \int \mathbf{M}_t(\alpha^*_{Q^{(N)}} - t)^2 \mathbf{Q}^{(N)}(dt) \qquad (10)$$

para cierta sucesión de distribuciones a priori $\mathbf{Q}^{(N)}$ (no obligatoriamente uniformes) y de estimaciones bayesianas correspondientes $\alpha^*_{Q^{(N)}}$, entonces α^* es una estimación minimax.

En nuestro caso, $m = \mathbf{M}_\alpha(\alpha_0^* - \alpha)^2$ no depende de α. Por eso, en virtud de las propiedades de convergencia anteriormente mencionadas,

$$\lim_{N\to\infty} \sup \int \mathbf{M}_t(\alpha^*_{Q^{(N)}} - t)^2 \mathbf{Q}^{(N)}\, dt) \geqslant$$

$$\geqslant \lim_{N\to\infty} \sup \frac{1}{2N} \int_{|t|<N-\sqrt{N}} \mathbf{M}_t(\alpha^*_{Q^{(N)}} - t)^2 dt \geqslant$$

$$\geqslant \lim_{N\to\infty} \sup \frac{1}{2N}\, 2(N - \sqrt{N})(m - \varepsilon) = m - \varepsilon$$

para cualquier $\varepsilon > 0$. Esto significa que se ha cumplido la propiedad (10).

Así pues, la *estimación de Pitman es minimax en la clase de todas las estimaciones del parámetro de desplazamiento* (el hecho de que ella sea minimax en la clase de estimaciones equivariantes, se desprende, evidentemente, de la eficacia).

Lo dicho también se puede interpretar del modo siguiente: la «peor» distribución a priori (véase el § 11) para el parámetro de desplazamiento es la distribución «uniforme en todo el eje».

Como indicación del carácter minimax de la estimación de Pitman también podría servir la dependencia (señalada más arriba) $\mathbf{M}_\alpha(\alpha_0^* - \alpha)^2$ de α (compárese con el teorema 11.2).

4. Acerca de las estimaciones óptimas del parámetro de escala. Como ya hemos indicado, el problema de estimación del parámetro de escala σ puede reducirse, en cierto sentido, al problema de estimación del parámetro de desplazamiento. Sea, por abreviar, $\mathscr{X} = (0, \infty) = \Theta$. En este caso, si $X \in \mathbf{P}_\sigma$, $\mathbf{P}_\sigma(A) = \mathbf{P}(A/\sigma)$, entonces $Y = \ln X = (\ln x_1, \ldots, \ln x_n) \in \mathbf{P}_\alpha^{(1)}$, donde $\alpha = \ln \sigma$, y la distribución $\mathbf{P}_\alpha^{(1)}$ tiene una densidad $y_1 = \ln x_1$ en el punto y $\left(\text{la condición } (A_\mu) \text{ se cumple, } \frac{d\mathbf{P}_1(x)}{d\mu} = f(x)\right)$, igual a (véase [11], pág. 53)

$$f\left(\frac{e^y}{\sigma}\right)\frac{e^y}{\sigma} = f(e^{y-\alpha})e^{y-\alpha} = f^{(1)}(y - \alpha),$$

$$f^{(1)}(y) = f(e^y)e^y.$$

Ahora bien, podemos apreciar muy bien el parámetro α con ayuda de la estimación de Pitman $\alpha^* = \alpha^*(Y)$, y luego suponer que $\sigma^*(X) = e^{\alpha^*(Y)}$. Es fácil notar que $\sigma^*(X)$ será equivariante, ya que

$$\sigma^*(cX) = e^{\alpha^*(Y+\ln c)} = e^{\alpha^*(Y)+\ln c} = c\sigma^*(X).$$

No obstante, aquí es importante señalar que la estimación de Pitman mini-

miza $\mathbf{M}_\alpha(\alpha^* - \alpha)^2$. Por lo tanto, la estimación σ^* obtenida minimizará

$$\mathbf{M}_\sigma\left(\ln \frac{\sigma^*}{\sigma}\right)^2 \tag{11}$$

y no la magnitud $\mathbf{M}_\sigma(\sigma^* - \sigma)^2$ de la cual se trataba generalmente. Pero en el problema de estimación equivariante del parámetro σ no era racional examinar la estimación estándar, puesto que ella, a distinción de (11), depende de la transformación de contracción aplicada simultáneamente a σ^* y σ. Aquí, como análogo de la estadística invariante S_0 servirá la estadística $(x_2/x_1, \ldots, x_n/x_1)$. A la par con (11) también es posible, naturalmente, examinar otros errores. Si, por ejemplo, minimizamos la magnitud

$$\mathbf{M}_\sigma\left(\frac{\sigma^*}{\sigma} - 1\right)^2,$$

entonces, la mejor estimación equivariante será

$$\sigma^* = \frac{\int \sigma^{-n-2} f(X/\sigma)\, d\sigma}{\int \sigma^{-n-3} f(X/\sigma)\, d\sigma} \tag{12}$$

(véase [33], p.).

Ejemplo 3. Detección de la fuente de radiación. Examinemos un ejemplo de un problema físico real, relacionado con las estimaciones de desplazamiento y escala.

Supongamos que en cierto punto desconocido z del espacio tridimensional se encuentra una fuente de radiación gamma. El problema consiste en determinar las coordenadas del punto z utilizando un detector plano (que coincide con uno de los planos de coordenada) y, fijando en este detector las trazas de radiación, o sea, las trazas de interacción de los cuantos gamma, emitidos por el punto z, con la superficie sensible del detector.

Este problema sería mucho más simple si tuviéramos una fuente de radiación de partículas cargadas de alta energía. Entonces podríamos poner, uno tras otro, dos detectores planos paralelos y fijar en ellos los puntos de paso (o sea, de interacción con la superficie de la pantalla) tan sólo de dos partículas. Esto nos daría las direcciones del vuelo de esas partículas y junto con ellas las coordenadas del punto z como punto de intersección de dichas direcciones. Sin embargo, para una radiación gamma poco intensa, que se utiliza en roentgenoscopia, esto es irrealizable y tan sólo se puede introducir un detector.

La dirección de propagación de los cuantos gamma emitidos es aleatoria y se distribuye uniformemente en la superficie de la esfera (si dicha dirección se determina por un punto en la esfera con centro en el punto z).

Para simplificar el problema examinemos su variante bidimensional. Supongamos que la fuente se encuentra en el plano de las variables (x, y),

en un punto desconocido $z = (\alpha, \sigma)$, $\sigma > 0$. El ángulo de dirección de la radiación, formado con el eje Oy, tiene una distribución uniforme en $[0, 2\pi]$. El detector sensible coincide con el eje de abscisas. Los resultados de las observaciones serán los puntos $x_1, x_2, \ldots$, en los que hemos fijado la interacción de los cuantos gamma con el detector (con el eje de abscisas).

La peculiaridad de este problema consiste en que el volumen n de la muestra obtenida durante un tiempo fijo t, será aleatorio: el número de cuantos gamma emitidos por la fuente en el tiempo t tiene una distribución de Poisson, y el número de cuantos gamma que alcanzaron el detector también está distribuido con arreglo a la ley de Poisson, ya que cada cuanto llega al eje de abscisas con una probabilidad igual a 1/2. No obstante, en nuestro caso, n y las observaciones $x_1, x_2, \ldots$ son independientes. Por eso podemos examinar el número n de observaciones que se ha obtenido y considerarlo fijo (para cada uno de tales números n fijos, la distribución de x_l será la misma).

Así pues, supongamos que se han dado las observaciones $X = (x_1, \ldots, x_n)$. Nuestro problema consiste en estimar las coordenadas (α, σ). Mostremos que $X \in K_{\alpha,\sigma}$, o sea, x_l tienen una distribución de Cauchy con parámetros de desplazamiento α y de escala σ.

En efecto, la distribución condicional del ángulo β entre la dirección del movimiento del cuanto gamma y el eje $(0, -y)$, a condición de que

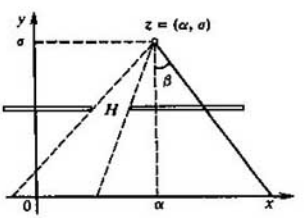

Fig. 2.

el cuanto haya alcanzado el detector (el eje de abscisas), será uniforme en el segmento $[-\pi/2, \pi/2]$. Como $(x - \alpha)/\sigma = \operatorname{tg}\beta$ (véase la fig. 2), entonces

$$\mathbf{P}_{\alpha,\sigma}(x_1 < x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} \frac{x - \alpha}{\sigma}.$$

Por consiguiente, la densidad de distribución de x_1 será igual a la densidad de distribución de Cauchy (véase el § 2)

$$k_{\alpha,\sigma}(x) = \frac{1}{\pi\sigma} \frac{1}{(1 + ((x - \alpha)/\sigma)^2)} = \frac{\sigma}{\pi(\sigma^2 + (x - \alpha)^2)}.$$

Ahora supongamos que σ es conocido, por ejemplo, $\sigma = 1$. Entonces la mejor estimación invariante del parámetro de desplazamiento α será la de Pitman, que se obtiene como el valor medio de $\alpha^* = \int u\varphi(u)du$ de la distribución con una densidad de

$$\varphi(u) = \varphi(u, X) = \frac{k_u(X)}{\int k_v(X)dv}, \quad k_u(X) = \prod_{i=1}^{n} k_u(x_i),$$

$$k_u(x_i) = k_{u,1}(x_i) = \frac{1}{\pi(1 + (u - x_i)^2)}.$$

La e.v.m $\hat{\alpha}^*$ será un punto en el que se alcanza el máx $\varphi(u)$. Más adelante mostraremos (véanse los §§ 24 y 25) que α^* y $\hat{\alpha}^*$ son asintóticamente equivalentes y tienen una distribución asintóticamente normal con coeficiente $1/I = 2$ (en el caso sujeto a examen $I = \int (k_0')^2/k_0 dx = 4\pi^{-1}\int x^2(1 + x^2)^{-3}dx = 1/2$). De lo dicho resulta que el error de las estimaciones α^* y $\hat{\alpha}^*$ para grandes n tiene un orden de pequeñez igual a $1/\sqrt{n}$.

Es interesante señalar que en el problema sometido a examen se puede alcanzar un grado más alto de exactitud, interviniendo en el experimento. Esto se puede hacer colocando entre el punto $z = (\alpha, 1)$ y el detector una pantalla paralela al eje de abscisas y provista del orificio H, a través del cual sólo pueden pasar los cuantos gamma. Las posiciones de la pantalla y el orificio se eligen según el experimentador y, por lo tanto, son conocidas.

En este caso la distribución de las observaciones en la pantalla será discontinua y, si los orificios H son pequeños, será próxima a $U_{a\alpha, a\alpha+b}$ para ciertas constantes a y b que conocemos. La forma de la estimación equivariante eficiente α_H^* para tal distribución fue hallado en el ejemplo 2. La estimación α_H^* se determina por los valores extremos de la muestra y tiene una exactitud del orden de $1/n_H$, donde $n_H \leqslant n$ es el número de elementos de la muestra, los cuales corresponden a los cuantos que han pasado a través de la ranura (n_H, al igual que n, es realmente aleatorio y está distribuido de acuerdo con la ley de Poisson). Como, por término medio, n_H es proporcional a n, con valores de n bastante grandes obtenemos $1/n_H \ll 1/\sqrt{n}$.

§ 19*. Problema general sobre la estimación equivariante

Examinemos el grupo G de transformaciones medibles g del espacio $\mathscr{X}^n$ en sí, que poseen las propiedades siguientes:

1) cada g aplica $\mathscr{X}^n$ en todo el espacio $\mathscr{X}^n$, o sea, para cada $x_2 \in \mathscr{X}^n$ se encontrará un $x_1 \in \mathscr{X}^n$ tal que $x_2 = gx_1$.

2) las aplicaciones g son biunívocas.

La mensurabilidad de g se necesita para que gX sea una variable aleatoria. La propiedad de grupo quiere decir que $g_2g_1 \in G$ si $g_1 \in G$, $g_2 \in G$; la transformación idéntica e y la inversa g^{-1} pertenecen a G (así que $g^{-1}g = e$).

Definición 1. La familia de distribuciones $\{\mathbf{P}_\theta\}$ se llama *equivariante respecto al grupo de transformaciones* G(o, para abreviar, simplemente invariante) si para cada $g \in G$ y $\theta \in \Theta$ existe el único $\theta_g \in \Theta$ tal que la relación $X \in \mathbf{P}_\theta$ conduce a $gX \in \mathbf{P}_{\theta_g}$.

Designemos por $\theta_g = \overline{g}\theta$ el valor de θ_g definible unívocamente por θ y g. Entonces la definición significa que

$$\mathbf{P}_\theta(gX \in A) = \mathbf{P}_{\overline{g}\theta}(X \in A).$$

Como en virtud de la definición 1 se cumple la condición (A_0), el conjunto $\overline{G}$ de todas las transformaciones $\overline{g}$ del espacio Θ en sí forma un grupo. En efecto, la distribución g_2g_1X se da simultáneamente por las distribuciones $\mathbf{P}_{\overline{g_2g_1}\theta}$ y $\mathbf{P}_{\overline{g}_2\overline{g}_1\theta}$. De la condición (A_0) resulta que $\overline{g_2g_1} = \overline{g}_2\overline{g}_1$ y que $g_1^{-1} \in G$ (es suficiente poner $g_2 = g_1^{-1}$). Las transformaciones $\overline{g}$ de $\overline{G}$ son automáticamente biunívocas. Sin embargo, puede no haber isomorfismo entre G y $\overline{G}$. Sea, por ejemplo, $X \in \Phi_{0,\sigma^2}$, $\sigma \in (0, \infty)$. En este caso la densidad $f_{0,\sigma^2}(X)$ (función de verosimilitud) depende exclusivamente de $\sum x_i^2$. Por consiguiente, si en calidad de G examinamos un grupo de revoluciones (transformaciones ortogonales de $\mathscr{L}^n$), entonces, las condiciones de la definición 1 serán cumplidas, pero $\overline{g} = \overline{e}$, y el grupo $\overline{G}$ se compone del único elemento $\overline{e}$, o sea, de la transformación idéntica de $\Theta = (0, \infty)$ en sí.

Le proponemos al lector que compruebe, en calidad de ejercicio, que si $\{\mathbf{P}_\theta\}$ es invariante respecto al grupo G, y G_1 es un subgrupo de G, entonces $\{\mathbf{P}_\theta\}$ es invariante respecto a G_1.

Cuando examinemos el problema general de estimación equivariante necesitaremos un planteamiento más general del problema respecto a la comparación de las estimaciones. Hasta ahora lo hemos hecho con ayuda de las desviaciones estándar, midiendo el error de la estimación por la magnitud $(\theta^* - \theta)^2$. Ahora supondremos que la medición del error de θ^* ocurre con ayuda de la función $w(\theta^*, \theta)$ y que esta función posee propiedad de "homogeneidad"*):

$$w(\overline{g}\theta, \overline{g}\theta^*) = w(\theta, \theta^*) \text{ para todos los valores de } \theta. \qquad (1)$$

Precisamente esta propiedad es típica de las funciones $w(\theta, \theta^*) = (\theta - \theta^*)^2$ para el parámetro de desplazamiento (transformación de desplazamiento)

*) Esta propiedad no es obligatoria en la teoría de estimación equivariante. Sólo se puede exigir la existencia de $\overline{g}\theta^*$ tal que para todos $\theta w(\overline{g}\theta, \overline{g}\theta^*) = w(\theta, \theta^*)$ (véase [33]).

y $w(\theta, \theta^*) = \left(\ln \frac{\theta}{\theta^*}\right)^2$ ó $\left(\frac{\theta}{\theta^*} - 1\right)^2$ para el parámetro de escala (transformación de contracción).

Hemos visto en el punto 4 del § 18, que el problema de determinación de la mejor estimación invariante puede ser muy sensible al elegir la medida del error $w(\theta, \theta^*)$ de la estimación θ^*.

Recurramos ahora al problema de estimación de las familias invariantes $\{\mathbf{P}_\theta\}$. Supongamos que tenemos la muestra X y que basándonos en ella hemos construido la estimación $\theta^* = \theta^*(X)$ del parámetro θ. Si examinamos la muestra $Y = gX \in \mathrm{P}_{\bar{g}\theta}$, entonces $\theta^*(Y)$ será la estimación para $\bar{g}\theta$. En este caso es natural suponer que las estimaciones $\theta^*(X)$ y $\theta^*(Y)$ están ligadas entre sí al igual que los parámetros sujetos a estimación θ y $\bar{g}\theta$, o sea, mediante la transformación $\bar{g}$:

$$\theta^*(Y) = \bar{g}\theta^*(X). \tag{2}$$

En virtud de (1), la estimación $\theta^*(Y)$ del parámetro $\bar{g}(\theta)$ proporciona el mismo error que la estimación $\theta^*(X)$ del parámetro θ. Por lo tanto, tenemos dos problemas de estimación "iguales". Las transformaciones realizadas gX y $\bar{g}\theta$ pueden interpretarse como las sustituciones de los sistemas de coordenadas. Entonces (2) significa que la estimación θ^* no depende de la elección del sistema de coordenadas y satisface la relación

$$\theta^*(X) = \bar{g}^{-1}\theta^*(gX). \tag{3}$$

Con otras palabras, si se ha elegido θ^*, que satisface (2), entonces no importa cuál de los dos problemas de estimación mencionados más arriba ha de ser resuelto, puesto que, mediante la igualdad (3), las deducciones acerca de $\bar{g}\theta$ en el segundo problema pueden convertirse en deducciones acerca de θ en el primer problema.

Definición 2. La estimación θ^* del parámetro θ de la familia invariante $\mathbf{P}_\theta$ que satisface (3) se llama *equivariante*[*)].

Examinemos cualquier punto $\theta_0 \in \Theta$ y el conjunto de puntos "equivalentes" $\theta = \bar{g}\theta_0$, $\bar{g} \in \bar{G}$. Tal formación de clases de puntos "equivalentes" divide todo el espacio Θ en subconjuntos llamados *órbitas*.

Teorema 1. *El valor de* $\mathbf{M}_\theta\, w(\theta, \theta^*)$ *para la estimación equivariante* θ^* *es constante en la órbita, o sea,*

$$\mathbf{M}_\theta w(\theta, \theta^*) = \mathbf{M}_{\bar{g}\theta} w(\bar{g}\theta, \theta^*)$$

para cualesquiera $\theta \in \Theta$ *y* $\bar{g} \in \bar{G}$.

[*)] Tales estimaciones se denominan, a veces, invariantes. Sin embargo, este término es menos exacto. Es mejor dejarlo para las estimaciones que poseen la propiedad $\theta^*(gX) = \theta^*(X)$ (o sea, para el caso cuando $\bar{g} = \bar{e}$ para todo g).

Demostración.

$$\mathbf{M}_\theta w(\theta,\ \theta^*(X)) = \mathbf{M}_\theta w(\bar{g}\theta,\ \bar{g}\theta^*(X)) =$$
$$= \mathbf{M}_\theta w(\bar{g}\theta,\ \theta^*(gX)) = \mathbf{M}_{\bar{g}\theta} w(\bar{g}\theta,\ \theta^*(X)). \quad \triangleleft$$

Si la órbita $\{\theta';\ \theta' = \bar{g}\theta_0,\ \bar{g} \in \overline{G}\}$ coincide con Θ (como tuvo lugar para los parámetros de desplazamiento y escala), entonces $\mathbf{M}_\theta w(\theta,\ \theta^*) = \text{const}$ en Θ. El cumplimiento de esta igualdad es el síntoma característico del carácter minimax de θ^* (compárese con el teorema 11.2), así que las mejores estimaciones equivariantes a menudo resultan minimax en la clase de todas las estimaciones (esto se detalla en [33]).

De los teoremas del § 11 se deduce, por ejemplo, el

Teorema 2. *Si Θ es una órbita, y la estimación equivariante θ^* resultó bayesiana (o el límite de estimaciones bayesianas θ_N^* con una convergencia $\mathbf{M}_\theta w(\theta,\ \theta^*) = \lim_{N\to\infty} \mathbf{M}_\theta w(\theta,\ \theta_N^*)$), entonces θ^* es una estimación minimax.*

Nótese también la siguiente propiedad importante de las estimaciones equivariantes. Será cómodo designar por $\nu(g,\ dx)/\nu(dx)$ la densidad de la medida $\nu_g,\ \nu_g(B) = \nu(gB)$ respecto a la medida ν en el punto $x \in \mathscr{X}^n$.

Teorema 3. *Supongamos que se cumple la condición (A_μ) y $\mu^n(g\ dx)/\mu^n(dx)$ es finito y positivo para cada $g \in G$, y c.t. $[\mu^n]$ valores de x. Supongamos, además, que la e.v.m. $\hat{\theta}^*$ es la única para cada X. En este caso, si la familia $\mathbf{P}_\theta$ es invariante, entonces $\hat{\theta}^*$ es la estimación equivariante.*

Demostración. Tenemos

$$f_{\hat{\theta}^*(X)}(X) = \frac{\mathbf{P}_{\hat{\theta}^*(X)}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_\theta(dx)}{\mu^n(dx)} \tag{4}$$

en el punto $x = X$. Suponiendo $Y = gX$, también podemos escribir

$$f_{\hat{\theta}^*(Y)}(Y) = \frac{\mathbf{P}_{\hat{\theta}^*(Y)}(g\ dx)}{\mu^n(g\ dx)} = \max_\theta \frac{\mathbf{P}_\theta(g\ dx)}{\mu^n(g\ dx)}.$$

En virtud de la invariación de $\mathbf{P}_\theta$ y del carácter finito de $\mu^n(g\ dx)/\mu^n(dx) > 0$, esto equivale a que

$$\frac{\mathbf{P}_{\bar{g}^{-1}\hat{\theta}(X)}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_{\bar{g}^{-1}}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_\theta(dx)}{\mu^n(dx)}.$$

Comparando con (4) y utilizando la unicidad de $\hat{\theta}^*(X)$, obtenemos $\bar{g}^{-1}\hat{\theta}^*(gX) = \hat{\theta}^*(X)$. $\triangleleft$

§ 20. Desigualdad integral del tipo Rao—Cramer. Criterios del carácter asintóticamente bayesiano y minimax de las estimaciones

Este párrafo también podría titularse "Desigualdad para la desviación estándar en el caso bayesiano". En su mayor parte el mismo se refiere a la teoría asintótica de la estimación.

Antes ya hemos tocado las cuestiones relacionadas con el enfoque asintótico de la comparación de las estimaciones. Ahora, y sobre todo en los §§ 23—29, dichas cuestiones serán el principal objeto de estudio.

1. Estimaciones eficientes y superefficientes. En el § 16, dedicado a la desigualdad de Rao—Cramer, quedó sin aclarar la siguiente cuestión importante. Supongamos que se cumple la cuestión (R). Entonces, para las estimaciones no desplazadas,

$$\mathbf{M}_\theta(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)}.$$

El segundo miembro de dicha desigualdad se llama, a veces, *frontera de Rao—Cramer.* Esta se alcanza para las estimaciones R-eficientes. La cuestión consiste en si ¿será posible o no, a costa de elegir el desplazamiento, mejorar considerablemente las estimaciones R-eficientes o asintóticamente R-eficientes? Es la cuestión acerca del carácter esencial de la frontera de Rao—Cramer y acerca del papel que desempeña el desplazamiento.

Ya hemos examinado parcialmente el hecho de que en un punto registrado θ_0, el valor de $\mathbf{M}_\theta(\theta^* - \theta)^2$ puede hacerse mucho menor que la frontera de Rao—Cramer. Para ello es suficiente tomar $\theta^* = \theta_0$. No obstante, en este caso, tal estimación en otros puntos será muy mala.

Se puede citar otro ejemplo menos trivial, donde el mejoramiento se alcanza no a expensas de otros puntos. Sea $X \in \Phi_{\alpha,1}$, $\alpha \in \Theta = [0, \infty)$. Entonces la estimación $\alpha^* = \bar{x}$ es eficiente e incluso R-eficiente. Sin embargo, en nuestro caso, cuando $\Theta = [0, \infty)$, la estimación $\alpha^{**} = \max(0, \bar{x})$ será, evidentemente, mejor, puesto que ella reduce las desviaciones estándar, sustituyendo por 0 los valores negativos inadmisibles. Es evidente que la estimación α^{**} ya será desplazada: $\mathbf{M}_\alpha\alpha^{**} > \alpha$, pero en el punto $\alpha = 0$ tenemos $I(\alpha) = 1$, $\mathbf{M}_0(\alpha^*)^2 = \frac{1}{n}$, $\mathbf{M}_0(\alpha^{**})^2 = \frac{1}{2n} < \frac{1}{nI(0)}$. En este ejemplo, el mejoramiento está relacionado con el hecho de que hemos reducido el campo de valores de la estimación α^* hasta el conjunto Θ. Citemos un ejemplo más (perteneciente a Hodges), en el que el mejoramiento de α^* ocurre no a costa de la limitación de Θ.

Sea, como antes, $X \in \Phi_{\alpha,1}$, $\alpha \in \Theta = (-\infty, \infty)$. Además de la estimación eficiente $\alpha^* = \bar{x}$ examinemos, cuando $\beta < 1$, la estimación

$$\alpha^{**} = \begin{cases} \bar{x} \text{ si } |\bar{x}| \geqslant n^{-1/4}, \\ \beta\bar{x} \text{ si } |\bar{x}| < n^{-1/4}. \end{cases}$$

No es difícil ver que, cuando $\alpha > 0$, según el teorema central, del límite,

$$\mathbf{P}_\alpha(|\bar{x}| < n^{-1/4}) \leqslant \mathbf{P}_\alpha((\bar{x} - \alpha)\sqrt{n} < n^{1/4} - \alpha\sqrt{n}) \to 0$$

cuando $n \to \infty$. La afirmación análoga es cierta cuando $\alpha < 0$. Por eso α^{**}, cuando $\alpha \neq 0$ α^{**}, coincide con $\bar{x}$ en el conjunto de la probabilidad que converge hacia 1 y, por lo tanto, según el teorema de continuidad cuando $\alpha \neq 0$,

$$(\alpha^{**} - \alpha)\sqrt{n} \in \Phi_{0,1}.$$

Cuando $\alpha = 0$,

$$\mathbf{P}_0(|\bar{x}| < n^{-1/4}) = \mathbf{P}_0(|\bar{x}\sqrt{n}| < n^{1/4}) \to 1$$

y α^{**} en el conjunto de la probabilidad convergente hacia 1 coincide con $\beta\bar{x}$, así que $(\alpha^{**} - \alpha)\sqrt{n} \in \Phi_{0,\beta^2}$. Por consiguiente, para todos los valores de α, la estimación α^{**} es asintóticamente normal, $(\alpha^{**} - \alpha)\sqrt{n} \in \in \Phi_{0,\sigma^2(\alpha)}$, donde

$$\sigma^2(\alpha) = \begin{cases} 1 & \text{cuando } \alpha \neq 0, \\ \beta^2 \quad 1 & \text{cuando } \alpha = 0. \end{cases}$$

Ahora bien, en el punto $\alpha = 0$, el coeficiente de dispersión $\sigma^2(0)$ resultó menor que la frontera inferior de Rao—Cramer, igual a 1.

Las estimaciones asintóticamente normales en los ejemplos citados, cuando el coeficiente de dispersión para ellas $\sigma^2(\theta) \leqslant I^{-1}(\theta)$ es, con algunos valores de θ, estrictamente menor que $I^{-1}(\theta)$, se llaman, a veces, *superficientes*.

No obstante, resultó que estos ejemplos cambian poco el cuadro, justo en general, acerca de la preferencia de las estimaciones eficientes. Precisamente Le Cam demostró que el mejoramiento (ilustrado más arriba) de las estimaciones, hablando en general, sólo se puede lograr en pequeñas cantidades de puntos.

En este párrafo mostraremos que a la par con la relación $\inf_{\theta^*} \mathbf{M}_t(\theta^* - t)^2 = 0$, válida para cada t, para la *integral* respecto a $\mathbf{M}_t(\theta^* - t)^2$ ya existe una *frontera inferior positiva que no depende de* θ^* y la cual se halla estrechamente relacionada con la integral análoga de la función $(nI(t))^{-1}$. Así mismo obtendremos, en el caso unidimensional $\theta \in R$, la desigualdad para

$$\inf_{\theta^*} \int \mathbf{M}_t(\theta^* - t)^2 q(t)dt, \tag{1}$$

cualquiera que sea la función ponderal $q(t) \geqslant 0$, $\int q(t)dt = 1$, cuyo segundo miembro no depende de θ^* (incluyendo también el desplazamiento $b(t)$ presente en la desigualdad de Rao—Cramer) y es próximo al valor de J/n, donde

$$J = \int \frac{q(t)}{I(t)}\, dt. \tag{2}$$

2. Desigualdades principales. Antes de enunciar los teoremas respectivos, señalaremos que la integral en (1) puede considerarse como la esperanza matemática incondicional $\mathbf{M}(\theta^* - \theta)^2$ en el caso bayesiano, cuando θ tiene distribución a priori, con una densidad $q(t)$ respecto a la medida de Lebesgue. En este caso $J = \mathbf{M}I^{-1}(\theta)$.

Designemos por $f(x, t) = f_t(x)\, q(t)$ la densidad de la distribución compatible de X, mientras que $\theta \cdot f_t'(x)$, como antes, designará la derivada de $f_t(x)$ respecto a t.

Seguidamente supongamos que $N_h \subset \Theta$ es el portador de la función h definida en Θ: $N_h = \{t: h(t) \neq 0\}$, y que N es el portador de $f(x, t)$ en $\mathscr{X}^n \times \Theta$.

Teorema 1. *Supongamos que* $f_t(x)$ *es derivable respecto a* t, *y que la función* $\sqrt{I(t)}$ *es integrable en cualquier intervalo finito. Entonces para toda función derivable* $h(t)$ *finita (o sea, igual a* 0 *fuera del intervalo finito), tal que* $N_h \subset N_q$, *es válida la desigualdad*

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \frac{[\mathbf{M}(h(\theta)/q(\theta))]^2}{n\mathbf{M}(I(\theta)[h(\theta)/q(\theta)]^2) + \mathbf{M}[h'(\theta)/q(\theta)]^2} = \frac{\left(\int h(t)dt\right)^2}{n\int I(t)h^2(t)/q(t)dt + \int (h'(t)^2)/q(t)dt}. \quad (3)$$

Demostración. Tenemos, en virtud del carácter finito de $h(t)$,

$$\int (f_t(x)h(t))'\,dt = \int d(f_t(x)h(t)) = 0,$$

$$\int t(f_t(x)h(t))'\,dt = -\int f_t(x)h(t)dt.$$

Por consiguiente, para toda θ^*,

$$\int_{\mathscr{X}^n} \int_{\Theta} (\theta^* - t)(f_t(x)\,h(t))'\,dt\mu^n(dx) = \int_{\mathscr{X}^n} \int_{\Theta} f_t(x)h(t)dt\mu^n(dx) = \int_{\Theta} h(t)dt. \quad (4)$$

Estas integrales pueden considerarse, en virtud de la condición $N_h \subset N_q$, como integrales respecto a N. Por lo tanto, podemos multiplicar y dividir por $f(x, t)$ la expresión subintegral en (4). Entonces obtenemos

$$\mathbf{M}\left[(\theta^* - \theta)\,\frac{(f_\theta(X)h(\theta))'}{f(X, \theta)}\right] = \int_{N_q} h(t)dt = \mathbf{M}\,\frac{h(\theta)}{q(\theta)}.$$

De aquí, en virtud de la desigualdad de Cauchy—Buniakovski, resulta

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \frac{[\mathbf{M}(h(\theta)/q(\theta))]^2}{\mathbf{M}[f_\theta(X)h(\theta))'/(f_\theta(X)q(\theta))]^2}. \quad (5)$$

Sólo queda reducir esta desigualdad a la forma (3). Nótese previamente que

$$\mathbf{M}_t|L'(X,\ t)| \leqslant n\sqrt{I(t)}$$

y que casi para todos*) t,

$$\mathbf{M}_t L'(X,\ t) = 0. \tag{6}$$

La primera de estas afirmaciones se deduce de las relaciones $\mathbf{M}_t|L'(X,\ t)| \leqslant n\mathbf{M}_t|l'(x_1 t)| \leqslant n\{\mathbf{M}_t[l'(x_1,\ t)]^2\}^{1/2} = n\sqrt{I(t)}$, que resulta de la desigualdad de Cauchy—Buniakovski. Para demostrar la segunda afirmación tomemos la función finita arbitraria $g(t)$ que en todas partes tiene la derivada continua $g'(t)$. Entonces

$$\int g(t)f_t(X)dt = -\int g'(t)f_t(X)dt.$$

Además,

$$\int |g(t)|\mathbf{M}_t|L'(X,\ t)|dt \leqslant n\int |g(t)|\sqrt{I(t)}\,dt < \infty.$$

De aquí resulta que se puede cambiar el orden de integración en la expresión siguiente:

$$\int g(t)\mathbf{M}_t L'(X,\ t)dt = \int_{\mathscr{X}^n}\int_{\Theta} g(t)f_t(x)dt\mu^n(dx) =$$

$$= -\int_{\mathscr{X}^n}\int_{\Theta} g'(t)f_t(x)dt\mu^n(dx) = -\int_{\Theta} g'(t)dt = -\int_{\Theta} dg(t) = 0.$$

El cumplimiento de esta igualdad para todos g precisamente significa la validez de (6).

Ahora podemos transformar el segundo miembro (5). Omitiendo, para abreviar, los argumentos de las funciones, obtenemos

$$\mathbf{M}\left[\frac{(f_\theta(X)h(\theta)'}{f_\theta(X)q(\theta)}\right]^2 = \mathbf{M}\left[L'\frac{h}{q} + \frac{h'}{q}\right]^2 = \mathbf{M}\left[\left(\frac{h}{q}\right)^2\mathbf{M}_\theta(L')^2\right] +$$

$$+ 2\mathbf{M}\left[\frac{h'h}{q^2}\mathbf{M}_\theta L'\right] + \mathbf{M}\left(\frac{h'}{q}\right)^2 = n\mathbf{M}\left[\left(\frac{h}{q}\right)^2 I\right] + \mathbf{M}\left(\frac{h'}{q}\right)^2.$$

Aquí hemos aprovechado el hecho de que, en virtud de (6),

$$\mathbf{M}\left[\frac{h'h}{q^2}\mathbf{M}_\theta L'\right] = \int_{\mathscr{X}_n}\frac{h'h}{q}\mathbf{M}_t L'\,dt = 0$$

y que (véase el § 16) $\mathbf{M}_\theta(L')^2 = nI(\theta)$. ◁

En las afirmaciones posteriores siempre supondremos que $f_t(x)$ satisface las condiciones del teorema 1.

*) En el § 16 hemos demostrado que esta igualdad, al cumplirse las condiciones (R), tiene lugar para todos t. Aquí nos será suficiente que la misma se cumpla para casi todos t.

Teorema 2. *Si la función* $h(t) = h_0(t) \equiv q(t)/I(t)$ *es finita y derivable, entonces*

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \left(\frac{J}{n}\left(1 + \frac{H}{nJ}\right)\right)^{-1} \geqslant \frac{J}{n} + \frac{H}{n^2}, \tag{7}$$

$$\text{donde } H = \int \left[\left(\frac{q(t)}{I(t)}\right)'\right]^2 \frac{dt}{q(t)}.$$

Observación 1. Las desigualdades dadas en los teoremas 1 y 2 son integrales desde el punto de vista de que pertenecen a las integrales de $\mathbf{M}_t(\theta^* - t)^2$. Desde este punto de vista las desigualdades del § 16 pueden llamarse *locales.*

Demostración. Esta afirmación se deduce directamente del teorema 1, ya que el segundo miembro en (3) se transforma, cuando $h = q/I$, en $J^2/(nJ + H)$. ◁

Por lo tanto, vemos que la frontera inferior de los posibles valores de $\mathbf{M}(\theta^* - \theta)^2$, con grandes valores de n, se distingue poco de la frontera $\frac{J}{n} = \int \frac{q(t)dt}{nI(t)}$ que es igual al valor de $\mathbf{M}(\theta_\delta^* - \theta)^2$ para la estimación R-eficiente θ_δ^*. Esto muestra que es racional utilizar las estimaciones eficientes, puesto que para ellas, cualquiera que sea la función q, casi se alcanza el valor extremal de $\mathbf{M}(\theta^* - \theta)^2$.

La estimación (7) es inmejorable, lo cual es confirmado por el

Ejemplo 1. Sea $X \in \Phi_{\alpha,1}$. Como sabemos, en este caso $I(\alpha) = 1$. Supongamos, luego, que el parámetro α se elige aleatoriamente con una densidad suave de $q(t)$, $t \in (-\infty, \infty)$. Entonces el segundo miembro de (7) se transforma en $(n + H)^{-1}$, donde

$$H = \int \frac{(q')^2}{q}\, dt = \mathbf{M}[(\ln q(\alpha))']^2.$$

Es nuestro caso, la estimación bayesiana α_δ^*, que corresponde a la distribución a priori $\mathbf{Q}$ con densidad q y que minimiza $\mathbf{M}(\alpha^* - \alpha)^2$, es igual a (véase el § 10)

$$\alpha_{\mathbf{Q}}^* = \frac{\int tq(t)f_t(X)dt}{\int q(t)f_t(X)dt} =$$

$$= \frac{\int tq(t)\exp(n\bar{x}t - t^2n/2)dt}{\int q(t)\exp(n\bar{x}t - t^2n/2)dt} = \frac{\int tq(t)\exp(-n(\bar{x} - t)^2/2)dt}{\int q(t)\exp(-n(\bar{x} - t)^2/2)dt}. \tag{8}$$

Es fácil hallar la representación asintótica de esta relación y mostrar que

$$\alpha_{\mathbf{Q}}^* = \bar{x} + \frac{q'(\bar{x})}{nq(\bar{x})} + O\left(\frac{1}{n^2}\right), \quad \mathbf{M}(\alpha_{\mathbf{Q}}^* - \alpha)^2 = \frac{1}{n} - \frac{H}{n^2} + O\left(\frac{1}{n^3}\right).$$

No obstante, procederemos más sencillamente, suponiendo que $q(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Entonces es evidente que $H = 1$, y el segundo miembro en (7) se convierte en $1/(n+1)$. Pero en el ejemplo 11.1 hemos establecido que

$$\mathbf{M}(\alpha_{\mathfrak{Q}}^* - \alpha)^2 = \frac{1}{n+1}.$$

De este modo, la inmejorabilidad de las igualdades (7) y (3) queda demostrada.

Teorema 3. *Si el intervalo* $(a - \varepsilon, a + \varepsilon)$ *se contiene en* Θ, *entonces, para toda estimación* θ^*,

$$\max_{t \in (a-\varepsilon, a+\varepsilon)} \mathbf{M}_t(\theta^* - t)^2 \geqslant \frac{1}{n \max\limits_{t \in (a-\varepsilon, a+\varepsilon)} I(t) + \pi^2 \varepsilon^{-2}}.$$

Demostración. Hagamos uso de la desigualdad

$$\max_{t \in (a-\varepsilon, a+\varepsilon)} \mathbf{M}_t(\theta^* - t)^2 \geqslant \int_{a-\varepsilon}^{a+\varepsilon} \mathbf{M}_t(\theta^* - t)^2 q(t)dt,$$

válida para toda densidad $q(t)$ que es igual a cero fuera de $(a - \varepsilon, a + \varepsilon)$. La afirmación necesaria se deduce del teorema 1 si suponemos en éste

$$h(t) = q(t) = \frac{1}{\varepsilon} \cos^2 \frac{\pi(t-a)}{2\varepsilon}, \quad |t - a| \leqslant \varepsilon.$$

Entonces

$$\mathbf{M}_t(\theta^* - \theta)^2 \geqslant \frac{1}{n \int I(t) q(t) dt + \int (q'(t))^2 / q(t) dt},$$

donde

$$\int \frac{(q'(t))^2}{q(t)} dt = \int_{-\varepsilon}^{\varepsilon} \frac{\left(\frac{\pi}{2\varepsilon^2} 2 \cos \frac{\pi t}{2\varepsilon} \operatorname{sen} \frac{\pi t}{2\varepsilon}\right)^2 \varepsilon}{\cos^2 \frac{\pi t}{2\varepsilon}} dt =$$

$$= \frac{1}{\varepsilon^2} \int_{-1}^{1} \pi^2 \operatorname{sen}^2 \frac{\pi t}{2} dt = \frac{\pi^2}{\varepsilon^2}. \quad \triangleleft$$

Se puede señalar que en la función $q(t) = \cos^2(\pi t/2)$ se alcanza el mínimo de la funcional $\int_{-1}^{1} (q'(t))^2/q(t)dt$ en la clase de todas las densidades derivables $q(t)$.

Del teorema 3 se deduce, en particular, que el intervalo de valores de θ para los cuales la estimación θ^* es supereficiente no puede tener una longitud mayor que $O(1/\sqrt{n})$.

3. Desigualdades en el caso cuando la función $q(\theta)/I(\theta)$ no es derivable. Si la función $h_0 = q/I$ no satisface las condiciones del teorema 1, es válida la siguiente afirmación útil que permite estimar la asintótica de $\mathbf{M}(\theta^* - \theta)^2$ en el caso general.

Teorema 4. *Supongamos que la sucesión de funciones $h_\varepsilon(t)$, dependientes del parámetro $\varepsilon > 0$, es tal que cada función h satisface las condiciones del teorema 1 y*

1) $h_\varepsilon(t) \leqslant h_0(t)$,

2) $H(\varepsilon) = \int \frac{(h_\varepsilon'(t))^2}{q(t)}\, dt < \infty$.

Entonces, para todo $\varepsilon > 0$,

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \frac{\left(\int h_\varepsilon(t)dt\right)^2}{nJ + H(\varepsilon)}.$$

La demostración se deduce directamente del teorema 1 si se toma $h = h_\varepsilon$.

Del teorema 4 obtenemos el siguiente colorario importante.

Teorema 5. *Si la función q es integrable según Riemann, $J < \infty$, entonces*

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \frac{J}{n}(1 + \delta_n),$$

donde $\delta_n = o(1)$ cuando $n \to \infty$.

Demostración. Pongamos $\hat{q}_\varepsilon(t) = \min\limits_{|u| \leqslant \varepsilon} q(t + u)$,

$$q_\varepsilon(t) = \begin{cases} \hat{q}_\varepsilon(t) & si\ \hat{q}_\varepsilon(t) \geqslant \varepsilon, \\ 0 & si\ \hat{q}_\varepsilon(t) < \varepsilon, \end{cases}$$

$$I_\varepsilon(t) = \text{máx}(\varepsilon,\ I(t)),$$

$$h_\varepsilon(t) = \frac{1}{2\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} \frac{q_\varepsilon(v)}{I_\varepsilon(v)}\, dv \leqslant h_0(t).$$

Es evidente que la función h_ε es finita y derivable para cualquier $\varepsilon > 0$.

Del hecho de que $q(t)$ es integrable según Riemann se desprende que $q_\varepsilon(t) \uparrow q(t)$ casi en todas las partes cuando $\varepsilon \to 0$. Para demostrar esto cerciorémonos de que

$$\int_a^b [q(t) - q_\varepsilon(t)]dt \downarrow 0. \tag{9}$$

De la integrabilidad de $q(t)$ según Riemann se deduce la convergencia

$$\sum_k q_\delta(2k\delta)2\delta \uparrow \int q(t)dt,$$

$$\sum_k q_\delta((2k + 1)\delta)2\delta \uparrow \int q(t)dt$$

cuando $\delta \to 0$. Por eso

$$\int_a^b q_\varepsilon(t)dt \geqslant \sum_k q_{2\varepsilon}(2k\varepsilon)2\varepsilon =$$

$$= \frac{1}{2}\left(\sum q_{2\varepsilon}(4k\varepsilon)4\varepsilon + \sum q_{2\varepsilon}((4k+2)\varepsilon)4\varepsilon\right) \to \int q(t)dt.$$

La relación (9), y junto con ella la convergencia de $q_\varepsilon(t) \uparrow q(t)$, quedan demostradas.

Utilizando ahora esta convergencia, obtenemos $\frac{q_\varepsilon(t)}{I_\varepsilon(t)} \uparrow h_0(t)$,

$$\int h_\varepsilon(t)dt = \int \frac{dt}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} \frac{q_\varepsilon(t+v)}{I_\varepsilon(t+v)}dv =$$

$$= \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} dv \int \frac{q_\varepsilon(t)}{I_\varepsilon(t)}\,dt = \int \frac{q_\varepsilon(t)}{I_\varepsilon(t)}\,dt \uparrow J.$$

Además,

$$|h'_\varepsilon(t)| = \frac{1}{2\varepsilon}\left|\frac{q_\varepsilon(t+\varepsilon)}{I_\varepsilon(t+\varepsilon)} - \frac{q_\varepsilon(t-\varepsilon)}{I_\varepsilon(t-\varepsilon)}\right| \leqslant \frac{q(t)}{\varepsilon^2},$$

$$H(\varepsilon) \leqslant \int \left(\frac{q(t)}{\varepsilon^2}\right)^2 q^{-1}(t)dt = \frac{1}{\varepsilon^4}.$$

Ahora podemos hacer uso del teorema 3. Suponiendo $\varepsilon = \varepsilon(n) = n^{-1/5}$, $n \to \infty$, obtenemos $\varepsilon(n) \to 0$,

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \frac{\left(\int h_\varepsilon(t)dt\right)^2}{nJ + n^{4/5}} = \frac{J}{n}(1 + o(1)). \quad \triangleleft$$

4. Algunos corolarios. Criterios del carácter asintóticamente bayesiano y minimax. Una de las principales conclusiones que pueden sacarse de los resultados de este párrafo consiste, hablando en general, en lo siguiente. Si existe la estimación asintóticamente R-eficiente, cualquiera que sea otra estimación que tomemos, no obtendremos "en total" (o "por término medio") un resultado asintóticamente mejor. Utilicemos este hecho, más tarde, en el § 25. Aquí sólo expondremos los criterios del carácter asintóticamente bayesiano y del carácter asintóticamente minimax que se desprenden directamente de los teoremas 2 y 5.

Definición 1. La estimación θ_1^*, que posee la propiedad

$$\mathbf{M}n(\theta_1^* - \theta)^2 = J + o(1) \tag{10}$$

cuando $n \to \infty$, se llama *R-bayesiana asintóticamente.*

Son las estimaciones para las cuales se alcanza asintóticamente la frontera inferior de las desviaciones estándar, definida en los teoremas 2.5. Las

mismas también podrían denominarse *estimaciones R-eficientes "en total"* (o "por término medio").

Recordemos (véase el § 11) que la estimación θ_1^* se llama *asintóticamente bayesiana* (con respecto a la distribución $\mathbf{Q}$) si para cualquier otra estimación θ^*

$$\lim_{n\to\infty} \sup [\mathbf{M}n(\theta_1^* - \theta)^2 - \mathbf{M}n(\theta^* - \theta)^2] \leqslant 0. \tag{11}$$

Corolario 1. *Supongamos que se cumplen las condiciones del teorema 1 y que la función $q(t)$ es integrable según Riemann. Entonces una estimación asintóticamente R-bayesiana es asintóticamente bayesiana.*

Demostración. Supongamos que θ_1^* es una estimación asintóticamente R-bayesiana. En virtud del teorema 5, para toda estimación θ^*,

$$\lim_{n\to\infty} \inf \mathbf{M}n(\theta^* - \theta)^2 \geqslant J.$$

De aquí y de (10) resulta (11).

También está claro que si existe una estimación asintóticamente R-bayesiana, toda estimación asintóticamente bayesiana será R-bayesiana (compárese con las observaciones referentes al teorema 16.3).

Del teorema 5 también se desprende el

Corolario 2. *Supongamos que se cumplen las condiciones del teorema 1 y que la función $q(t)$ es integrable según Riemann. Si θ_1^* y θ_2^* son dos estimaciones asintóticamente R-bayesianas, éstas son asintóticamente equivalentes desde el punto de vista siguiente:*

$$\mathbf{M}n(\theta_1^* - \theta_2^*)^2 \to 0, \ (\theta_1^* - \theta_2^*)\sqrt{n} \underset{P}{\to} 0,$$

donde la convergencia en probabilidad se entiende respecto a la distribución compatible de X y θ en $\mathscr{X}^n \times \Theta$.

La demostración es completamente análoga a las demostraciones de los teoremas 8.2, 16.4. La igualdad inicial (8.11), en virtud del teorema 5, da

$$\lim_{n\to\infty} \sup \mathbf{M}n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \quad \lhd$$

En los §§ 8 y 11 hemos señalado que para comparar las estimaciones, a la par con los valores *medios* $\int q(t)\mathbf{M}_t(\theta^* - t)^2 dt$, pueden considerarse los valores máximos

$$\sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2, \ \Gamma \subset \Theta.$$

En calidad de Γ se toma todo el conjunto Θ o la parte de éste que, según datos previos, contiene el valor desconocido de θ. Recordemos que la estimación $\bar{\theta}^*$ se llama minimax cuando para toda estimación θ^*

$$\sup_{t\in\Gamma} \mathbf{M}_t(\bar{\theta}^* - t)^2 \leqslant \sup_{t\in\Gamma} \mathbf{M}_t(\theta^* - t)^2.$$

La estimación θ_1^* se llama *asintóticamente minimax* cuando para toda estimación θ^*

$$\lim_{n\to\infty} \sup_{t\in\Gamma} \mathbf{M}_t[\sqrt{n}(\theta_1^* - t)]^2 \leqslant \lim_{n\to\infty} \inf \sup_{t\in\Gamma} \mathbf{M}_t[\sqrt{n}(\theta^* - t)]^2.$$

Corolario 3. *Supongamos que la información de Fisher $I(\theta)$ existe y es continua. En este caso, si para cualquier segmento $\Gamma \subset \Theta$,*

$$\lim_{n\to\infty} \sup_{t\in\Gamma} \mathbf{M}_t[\sqrt{n}(\theta_1^* - t)]^2 \leqslant \sup_{t\in\Gamma} I^{-1}(t), \tag{12}$$

entonces la estimación θ_1^* es asintóticamente minimax.

Demostración. Es suficiente convencerse de que para cualquier estimación θ^*,

$$\lim_{n\to\infty} \inf \sup_{t\in\Gamma} \mathbf{M}_t[\sqrt{n}(\theta^* - t)]^2 \geqslant \sup_{t\in\Gamma} I^{-1}(t). \tag{13}$$

Para cualquier distribución $\mathbf{Q}$ en Γ, con una densidad suave $q(t)$ respecto a la medida de Lebesgue,

$$\sup_{t\in\Gamma} \mathbf{M}_t[\sqrt{n}(\theta^* - t)]^2 \geqslant \int \mathbf{M}_t[\sqrt{n}(\theta^* - t)]^2 q(t)dt.$$

Según el teorema 2, la integral del segundo miembro es para cualquier estimación θ^*, no menor que $J - H/n$. Por eso el primer miembro de (13) es mayor o igual a

$$J = \int I^{-1}(t)q(t)dt.$$

Pero q es una densidad suave arbitraria y, para un valor dado de $\varepsilon > 0$, la misma siempre se puede elegir, en virtud de la continuidad de $I^{-1}(t)$, de modo que

$$J \geqslant \sup_{t\in\Gamma} I^{-1}(t) - \varepsilon.$$

En vista de que ε es arbitrario, (13) queda demostrada. ◁

En conclusión de este apartado es necesario hacer una observación importante, que consiste en que, al buscar las estimaciones asintóticamente óptimas, es posible limitarse a la clase $\bar{K}_0$ de estimaciones asintóticamente no desplazadas, que hemos introducido en el § 16. Esto se deduce de las consideraciones siguientes.

Ya hemos señalado que el segundo miembro de la desigualdad del teorema 5, equivalente a $J/n + o(1/n)$, no depende absolutamente del desplazamiento $b(\theta)$. Al mismo tiempo, si al construir la frontera inferior de $\mathbf{M}(\theta^* - \theta)^2$ partimos de la desigualdad de Rao—Cramer dada en el § 16, entonces obtendremos

$$\mathbf{M}(\theta^* - \theta)^2 \geqslant \min_b \int q(t)\left[\frac{(1 + b'(t))^2}{nI(t)} + b^2(t)\right]dt.$$

Se puede mostrar (compárese con [47]) que este valor mínimo de todos los desplazamientos $b(\theta)$ tiene (con ciertas suposiciones acerca de la suavidad de $q(t)$ y $I(t)$) esa misma forma $J/n + o(1/n)$ y (lo cual es esencial para nosotros) se alcanza en el desplazamiento $b(\theta)$ que posee, cuando $n \to \infty$, las propiedades

$$b'(t) = o(1),\ b(t) = o(1/\sqrt{n}).$$

La clase de estimaciones θ^* con tales desplazamientos es precisamente $\bar{K}_0$ (véase el § 16). La salida de θ^* de la clase $\bar{K}_0$ hace inaccesible la frontera $J/n + o(1/n)$. Ahora bien, en el enfoque asintótico, cuando las estimaciones asintóticamente normales se comparan con ayuda de los valores de $\mathbf{M}(\theta^* - \theta)^2$ cuando son suaves $q(t)$ e $I(t)$, es posible limitarse a examinar las estimaciones de la clase $K = K_{\Phi,2} \cap \bar{K}_0$ (hemos examinado la clase $K_{\Phi,2}$ en el § 8), puesto que las estimaciones fuera de la clase $\bar{K}_0$ son "inadmisibles" desde el punto de vista antes indicado.

5. Caso multidimensional. En el caso de $\theta \in R^k$ se pueden obtener los análogos para todos los teoremas de este párrafo y hacer las mismas deducciones que hemos obtenido para el caso unidimensional.

En particular, la afirmación del teorema 5, uno de los principales en este apartado, tendrá la forma

$$d^2 \geqslant J/n + o(1/n),$$

donde $d^2 = \|d_{ij}\|$, $d_{ij} = \mathbf{M}(\theta_i^* - \theta_i)(\theta_j^* - \theta_j)$, $J = \mathbf{M}I^{-1}(\theta)$.

Los razonamientos relacionados con las estimaciones bayesianas y minimax también conservan su validez cuando en calidad de error de la estimación se considera el valor

$$v(\theta^*) = \mathbf{M}_\theta(\theta^* - \theta)V(\theta^* - \theta)^T,$$

donde V es una matriz definida no negativamente. Deben llamarse estimaciones bayesianas o minimax (o asintóticamente bayesianas y minimax) las estimaciones cuyos errores satisfacen las desigualdades respectivas para cualquier matriz V definida no negativamente.

§ 21. Distancias de Kullback—Leibler, de Hellinger y χ^2. Sus propiedades

Los resultados de este párrafo serán esenciales para la obtención de los resultados principales de la teoría asintótica de estimación, así como para los resultados del cap. 3.

1. Definiciones y propiedades principales de las distancias.

Sean $\mathbf{P}$ y $\mathbf{G}$ dos distribuciones en $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ absolutamente continuas respecto a la medida μ. Designemos

$$\frac{d\mathbf{P}}{d\mu} = p,\ \frac{dG}{d\mu} = g,$$

N_p es el portador de la distribución $\mathbf{P}$: $N_p = \{x\colon p(x) > 0\}$.

Definición 1. Se llama *distancia de Kullback—Leibler* entre las distribuciones **P** y **G** la magnitud

$$\varrho_1(\mathbf{P}, \mathbf{G}) = \int_{N_p} \ln \frac{p(x)}{g(x)}\, \mathbf{P}(dx) = \int_{N_p} \ln \frac{p(x)}{g(x)}\, p(x)\mu(dx).$$

De hecho $\varrho_1(\mathbf{P}, \mathbf{G})$ no es, por supuesto, una distancia o una métrica en sentido general, ya que $\varrho_1(\mathbf{P}, \mathbf{G})$ no es una función simétrica de **P** y **G**. No obstante, veremos que $\varrho_1(\mathbf{P}, \mathbf{G})$ caracteriza en realidad (desde el punto de vista estadístico) la desviación de **G** respecto a **P**.

De la desigualdad $\ln(1 + v) - v \leqslant 0$ y la representación

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int \left[\ln \frac{g}{p} - \left(\frac{g}{p} - 1\right)\right] p\mu(dx)$$

se deduce que siempre $\varrho_1(\mathbf{P}, \mathbf{G}) \geqslant 0$. En el lema 6.1 hemos establecido que la desigualdad $\varrho_1(\mathbf{P}, \mathbf{G}) = 0$ sólo es posible si $\mathbf{P} = \mathbf{G}$.

Definición 2. Llamaremos *distancia* χ^2 entre las distribuciones **P** y **G** la magnitud

$$\varrho_2(\mathbf{P}, \mathbf{G}) = \int_{N_p \cup N_g} \frac{(p(x) - g(x))^2}{p(x)}\, \mu(dx).$$

Casi todas las observaciones hechas para la definición 1 se refieren a esta distancia. La denominación de χ^2 se explica por razones que serán aclaradas más tarde.

Definición 3. Se llama *distancia de Hellinger* entre las distribuciones **P** y **G** la magnitud

$$\varrho_3(\mathbf{P}, \mathbf{G}) = \int_{N_p \cup N_g} \left(\sqrt{p(x)} - \sqrt{g(x)}\right)^2 \mu(dx).$$

La distancia de Hellinger ya es la función simétrica de **P** y **G**, y el valor de $\sqrt{\varrho_3(\mathbf{P}, \mathbf{G})}$ posee todas las propiedades de la métrica (entre las funciones $\sqrt{p(x)}$ y $\sqrt{g(x)}$ en el espacio métrico $L_2(\mathscr{X}, \mu)$). Es fácil notar que

$$\varrho_3(\mathbf{P}, \mathbf{G}) = 2\left(1 - \int \sqrt{pq}\, \mu(dx)\right) \leqslant 2. \tag{1}$$

Las tres distancias introducidas desempeñan un papel importante en distintos problemas de la estadística matemática. Nos convenceremos de ello en cierta medida.

Si mediante estas distancias se caracteriza el grado de proximidad de las distribuciones, cuando la relación p/g es próxima a 1, resultará que todas ellas se comportan asintóticamente igual, con una exactitud de hasta los factores constantes. En efecto, valiéndose del desarrollo

$$\ln \frac{g}{p} = \ln\left(1 + \left(\frac{g}{p} - 1\right)\right) = \left(\frac{g}{p} - 1\right) - \frac{1}{2}\left(\frac{g}{p} - 1\right)^2 + O\left(\left|\frac{g}{p} - 1\right|\right)^3,$$

obtenemos

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int \ln\frac{g}{p}\cdot p\mu(dx) \leqslant \frac{1}{2}\int\left(\frac{g}{p}-1\right)^2 p\mu(dx) = \frac{1}{2}\varrho_2(\mathbf{P}, \mathbf{G}),$$

$$\varrho_2(\mathbf{P}, \mathbf{G}) = \int\frac{(p-g)^2}{p}\mu(dx) = \int(\sqrt{p}-\sqrt{g})^2\left(1+\sqrt{\frac{g}{p}}\right)^2\mu(dx) \leqslant$$

$$\leqslant 4\varrho_3(\mathbf{P}, \mathbf{G}).$$

De la última igualdad también se deduce que $\varrho_2(\mathbf{P}, \mathbf{G}) \geqslant \varrho_3(\mathbf{P}, \mathbf{G})$.

Además, $\varrho_1(\mathbf{P}, \mathbf{G}) \geqslant \varrho_3(\mathbf{P}, \mathbf{G})$. En efecto, como $\ln(1+x) \leqslant x$, entonces

$$\ln\frac{g}{p} = 2\ln + \left(1+\left(\sqrt{\frac{g}{p}}-1\right)\right) \leqslant 2\left(\sqrt{\frac{g}{p}}-1\right),$$

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int\ln\frac{g}{p}\,p\mu(dx) \geqslant -2\left(\int\sqrt{pg}\,\mu(dx)+1\right) = \varrho_3(\mathbf{P}, \mathbf{G}).$$

En lo sucesivo examinaremos el caso paramétrico y consideraremos que se cumple la condición $(A\mu)$. Nos interesarán las distancias ϱ_i, $i = 1, 2, 3$, entre las distribuciones $\mathbf{P} = \mathbf{P}_{\theta_1}$ y $\mathbf{G} = \mathbf{P}_{\theta_2}$ en $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$, así como entre las distribuciones muestrales correspondientes (aquí las designaremos por $\mathbf{P}^n_{\theta_1}$, $\mathbf{P}^n_{\theta_2}$) en $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$. (Señalemos que las distancias tienen sentido para las distribuciones arbitrarias, y con la naturaleza de los espacios no están relacionadas de ningún modo). Si $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$, podemos escribir

$$\varrho_1(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \int\ln\frac{f_{\theta_1}}{f_{\theta_2}}f_{\theta_1}\mu(dx) = \mathbf{M}_{\theta_1}\ln\frac{f_{\theta_1}(\mathrm{x}_1)}{f_{\theta_2}(\mathrm{x}_1)},$$

$$\varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \int\frac{(f_{\theta_1}-f_{\theta_2})^2}{f_{\theta_1}}\mu(dx) = \mathbf{M}_{\theta_1}\left(\frac{f_{\theta_2}(\mathrm{x}_1)}{f_{\theta_1}(\mathrm{x}_1)}-1\right)^2, \qquad (2)$$

$$\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \int\left(\sqrt{f_{\theta_1}}-\sqrt{f_{\theta_2}}\right)^2\mu(dx) = \mathbf{M}_{\theta_1}\left(\sqrt{\frac{f_{\theta_2}(\mathrm{x}_1)}{f_{\theta_1}(\mathrm{x}_1)}}-1\right)^2.$$

Si no se cumple la condición $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$, entonces $\varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$, $\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$ serán mayores que las esperanzas matemáticas correspondientes en (2).

Cabe señalar que a la par con (2) tiene lugar la siguiente igualdad útil que se desprende de (1):

$$\mathbf{M}_{\theta_1}\sqrt{f_{\theta_2}(\mathrm{x}_1)/f_{\theta_1}(\mathrm{x}_1)} = \int\sqrt{f_{\theta_2}(x)f_{\theta_1}(x)}\,\mu(dx) =$$

$$= 1-\frac{1}{2}\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}). \qquad (3)$$

La relación entre las distancias $\varrho_i(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$ y $\varrho_i(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2})$ se establece por la afirmación siguiente.

Teorema 1.

$$\varrho_1(\mathbf{P}_{\theta_1}^n, \mathbf{P}_{\theta_2}^n) = n\varrho_1(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}),$$
$$1 + \varrho_2(\mathbf{P}_{\theta_1}^n, \mathbf{P}_{\theta_2}^n) = (1 + \varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}^n))^n, \tag{4}$$
$$1 - \frac{1}{2}\varrho_3(\mathbf{P}_{\theta_1}^n, \mathbf{P}_{\theta_2}^n) = \left(1 - \frac{1}{2}\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})\right)^n.$$

La demostración es casi evidente si se supone, para abreviar, que $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$ (en el caso general los cálculos conservarán, de hecho, su validez, pero serán un poco más voluminosos). En efecto, en este caso podemos hacer uso de las igualdades (2). Entonces la primera de las relaciones (4) se deduce directamente del hecho de que

$$\ln \frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} = \sum_{i=1}^{n} \ln \frac{f_{\theta_1}(x_i)}{f_{\theta_2}(x_i)}.$$

Seguidamente, en virtud de (2),

$$1 + \varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \mathbf{M}_{\theta_1}(f_{\theta_2}(x_1)/f_{\theta_1}(x_1))^2,$$
$$1 - \varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})/2 = \mathbf{M}_{\theta_1}\sqrt{f_{\theta_2}(x_1)/f_{\theta_1}(x_1)},$$

y las relaciones de este mismo tipo son válidas para las distancias entre $\mathbf{P}_{\theta_1}^n$ y $\mathbf{P}_{\theta_2}^n$ (sustituyendo en los segundos miembros x_1 por X). Como

$$\mathbf{M}_{\theta_1}\left(\frac{f_{\theta_2}(X)}{f_{\theta_1}(X)}\right)^{\alpha} = \mathbf{M}_{\theta_1}\prod_{i=1}^{n}\left(\frac{f_{\theta_2}(x_i)}{f_{\theta_1}(x_i)}\right)^{\alpha} = \left[\mathbf{M}_{\theta_1}\left(\frac{f_{\theta_2}(x_1)}{f_{\theta_1}(x_1)}\right)^{\alpha}\right]^n,$$

de aquí, cuando $\alpha = 2$ y $\alpha = 1/2$, obtenemos (4).

Le recomendamos al lector que demuestre este teorema en el caso general (o sea, cuando no se cumple la condición $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$). ◁

Del teorema 1 se desprende el

Corolario 1.

$$\varrho_3(\mathbf{P}_{\theta_1}^n, \mathbf{P}_{\theta_2}^n) \leqslant n\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}).$$

En efecto, $1 - \beta^n \leqslant (1 - \beta)n$ para cualquier $\beta \geqslant 0$. Suponiendo $\beta = 1 - \frac{1}{2}\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$, obtenemos de (4),

$$\varrho_3(\mathbf{P}_{\theta_1}^n, \mathbf{P}_{\theta_2}^n) = 2(1 - \beta^n) \leqslant 2(1 - \beta)n = n\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}). \quad \triangleleft$$

2. Relación de las distancias de Hellinger y otras con la información de Fisher. Entre las tres distancias introducidas en el apartado anterior, en lo sucesivo, la distancia de Hellinger tendrá para nosotros, el mayor interés. Al mismo tiempo, el carácter de las afirmaciones principales, expuestas más abajo (teoremas 2 y 3), y el carácter de las demostraciones serán iguales para las tres distancias. Por eso, para abreviar, nos limitare-

mos, en este apartado, a estudiar la distancia de Hellinger, que designaremos (omitiendo el índice del símbolo ϱ_3) del siguiente modo:

$$\varrho(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \int \left(\sqrt{f_{\theta_1}} - \sqrt{f_{\theta_2}}\right)^2 \mu(dx).$$

Pongamos $r(\theta_1, \theta_2) = \varrho(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$.

Lema 1. *Si $f_\theta(x)$, para c.t. $[\mu]$ valores de x, es continua respecto a θ, $\theta_1 \neq \theta_2$, entonces*

$$\lim_{\substack{\theta' \to \theta_1 \\ \theta'' \to \theta_2}} \inf \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2}. \tag{5}$$

Si la función $\sqrt{f_\theta(x)}$, para c.t. $[\mu]$ valores de x, es derivable respecto a θ, entonces

$$\lim_{\substack{\theta' \to \theta \\ \theta'' \to \theta}} \inf \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \frac{I(\theta)}{4}. \tag{6}$$

Además,

$$\frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant \frac{1}{4} \int_0^1 I(\theta_1 + (\theta_2 - \theta_1)y)dy. \tag{7}$$

Aquí se supone, claro está, que los valores de θ', θ'', θ_1, θ_2, θ pertenecen a Θ.

Demostración. Para verificar (5) es suficiente utilizar el lema de Fatou y la continuidad de $f_\theta(x)$ en la relación

$$\lim_{\substack{\theta' \to \theta_1 \\ \theta'' \to \theta_2}} \inf \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \int \lim_{\substack{\theta' \to \theta_1 \\ \theta'' \to \theta_2}} \inf \left(\frac{\sqrt{f_{\theta'}} - \sqrt{f_{\theta''}}}{\theta' - \theta''}\right)^2 \mu(dx).$$

En vista de que, cuando $\theta_1 = \theta_2 = \theta$, la expresión subintegral en la última integral es igual a $(f'_\theta)^2/(4f_\theta)$, obtenemos (6).

Para demostrar (7) pongamos $a = \theta_2 - \theta_1$ y representemos el incremento $\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}}$ en la forma

$$\frac{1}{2} \int_{\theta_1}^{\theta_2} \frac{f'_t}{\sqrt{f_t}}\, dt = \frac{a}{2} \int_0^1 \frac{f'_{\theta_1 + ay}}{\sqrt{f_{\theta_1 + ay}}}\, dy.$$

En virtud de la desigualdad de Cauchy—Buniakovski,

$$(\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}})^2 = \left[\frac{a^2}{4} \int_0^1 \frac{f'_{\theta_1 + ay}}{\sqrt{f_{\theta_1 + ay}}}\, dy\right]^2 \leqslant \frac{a^2}{4} \int_0^1 \frac{(f'_{\theta_1 + ay})^2}{f_{\theta_1 + ay}}\, dy.$$

Utilizando la negatividad de la función subintegral, podemos cambiar el

orden de integración en las relaciones siguientes:

$$\frac{r(\theta_1,\ \theta_2)}{a^2} \leqslant \frac{1}{4} \int\limits_{\mathscr{X}} \left(\int\limits_0^1 \frac{\left(f'_{\theta_1+ay}\right)^2}{f_{\theta_1+ay}}\, dy \right) \mu(dx) = \frac{1}{4} \int\limits_0^1 I(\theta_1 + ay)dy.$$

La desigualdad (7) queda demostrada. ◁

Pongamos $r(\Delta) = r(\theta,\ \theta + \Delta)$. Del lema 1 se deduce directamente el

Teorema 2. *Si la función* $\sqrt{f_\theta(x)}$, *para c.t.* $[\mu]$ *valores de x, es derivable respecto a* θ, *e* $I(\theta)$ *es continua, entonces existe*

$$\lim_{\Delta \to 0} \frac{r(\Delta)}{\Delta^2} = \frac{I(\theta)}{4}. \tag{8}$$

Observación 1. Esta afirmación también será válida para las distancias ϱ_1 y ϱ_2 si suponemos

$$r(\Delta) = \frac{1}{4}\, \varrho_2(\mathbf{P}_\theta,\ \mathbf{P}_{\theta+\Delta}),\ r(\Delta) = \frac{1}{2}\, \varrho_1(\mathbf{P}_\theta,\ \mathbf{P}_{\theta+\Delta}).$$

En este caso, la relación (6) se demuestra exactamente igual que en el lema 1. La demostración de (8) puede exigir la utilización de condiciones adicionales de regularidad (próximas a las condiciones (R)) que aseguren la validez del paso límite bajo el signo integral.

Así pues, $\varrho_i(\mathbf{P}_\theta,\ \mathbf{P}_{\theta+\Delta})$, $i = 1, 2, 3$, se comportan asintóticamente igual, e $I(\theta)$ caractariza la velocidad de su tendencia hacia el cero cuando $\Delta \to 0$ $\left(\text{pues } \frac{1}{4}\, I(\theta) \text{ es la segunda derivada de } r(v) \text{ en el punto } v = 0\right)$.

Si se pone $r^{(n)}(\Delta) = \varrho(\mathbf{P}^n_{\theta+\Delta},\ \mathbf{P}^n_\theta)$, entonces, de los teoremas 1 y 2 resultará

$$\lim_{\Delta \to 0} \frac{r^{(n)}(\Delta)}{\Delta^2} = \frac{nI(\theta)}{4}.$$

Estas mismas relaciones se mantendrán para las distancias ϱ_1 y ϱ_2.

3. Existencia de fronteras uniformes para $r(\Delta)/\Delta^2$. En lo sucesivo, la existencia de tales fronteras nos permitirá obtener estimaciones muy útiles para los momentos de relación de verosimilitud.

A fin de simplificar la exposición o evitar la introducción de otras condiciones más voluminosas, en las investigaciones posteriores a menudo estimaremos que se cumple la condición

(A_c): *el conjunto* Θ *es compacto.*

Desde el punto de vista de las aplicaciones, esta condición, que significa el carácter limitado y cerrado del conjunto paramétrico, por lo general, no es limitativa.

Más adelante también utilizaremos la condición (A_0) que hemos introducido en el § 6 y que significa que $f_{\theta_1} \neq f_{\theta_2}$ cuando $\theta_1 \neq \theta_2$. Con esta condición, $r(\theta_1,\ \theta_2) > 0$ cuando $\theta_1 \neq \theta_2$.

Teorema 3. *Si se cumplen las condiciones* (A_0), (A_c), *y* $0 < I(\theta) \leqslant 4h < \infty$ *para todos* $\theta \in \Theta$, *entonces existe una constante* $g > 0$ *tal que para todos* $\theta_1, \theta_2 \in \Theta$,

$$g \leqslant \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant h. \tag{9}$$

Demostración. La estimación superior se deduce directamente de (7). Mostremos ahora, que

$$\inf_{\theta_1, \theta_2} \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \geqslant g > 0. \tag{10}$$

Supongamos que (10) no es cierta, entonces habrá una sucesión $\theta_1^{(n)}$, $\theta_2^{(n)}$ tal que

$$\frac{r(\theta_1^{(n)}, \theta_2^{(n)})}{|\theta_1^{(n)} - \theta_2^{(n)}|^2} \to 0 \tag{11}$$

cuando $n \to \infty$. En virtud de la condición (A_c) podemos considerar, sin limitar la generalidad, que $\theta_1^{(n)} \to \theta_1 \in \Theta$, $\theta_2^{(n)} \to \theta_2 \in \Theta$. Si $\theta_1 \neq \theta_2$, entonces (11) contradice (5), ya que, en virtud de la condición (A_0), $r(\theta_1, \theta_2) > 0$. Pero si $\theta_1 = \theta_2 = \theta$, entonces (11) contradice (6), ya que $I(\theta) > 0$. El teorema queda demostrado.

4. Caso multidimensional. En este apartado obtendremos los análogos de las afirmaciones de los puntos 2 y 3 para el parámetro multidimensional (el contenido del punto 1 no está relacionado con la dimensión de θ). Designemos por $\varphi(x, \theta)$ la función vectorial con coordenadas

$$\varphi_i(x, \theta) = \frac{1}{\sqrt{f_\theta(x)}} \frac{\partial f_\theta(x)}{\partial \theta_i}.$$

Entonces la derivada de la función $\sqrt{f_\theta(x)}$ en el sentido del vector unitario $\omega = (\omega_1, \ldots, \omega_k)$ es igual a $((\sqrt{f_\theta(x)})', \omega) = (\operatorname{grad} \sqrt{f_\theta(x)}, \omega) = \frac{1}{2}(\varphi(x, \theta), \omega)$. La matriz de Fisher $I(\theta)$ en estas designaciones es igual a

$$I(\theta) = \int \varphi^T(x, \theta), \varphi(x, \theta) \mu(dx).$$

Supongamos que $|u|$ significa la norma euclídea $u = (u_1, \ldots, u_k)$.

En el caso multidimensional tiene lugar la siguiente generalización del lema 1.

Lema 1A. *La primera afirmación del lema 1* (véase (5)) *conserva por completo su validez cuando* $k > 1$.

Si la función $\sqrt{f_\theta(x)}$, *cuando c.t.* $[\mu]$ *valores de x, es derivable respecto a* θ, $\theta' \to \theta$, $\theta'' = \theta' + \omega'' \delta$, $\omega'' \to \omega$, $|\omega''| = |\omega| 1$, $\delta \to 0$, *entonces*

$$\liminf \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \frac{1}{4} \omega I(\theta) \omega^T. \tag{12}$$

Además, si ω, $|\omega| = 1$ *es un vector colineal a* $\theta_2 - \theta_1$, *de modo que* $\theta_2 = \theta_1 + a\omega$, $a = |\theta_2 - \theta_1|$, *entonces*

$$\frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant \frac{1}{4} \int_0^1 \omega I(\theta_1 + a\omega y)\omega^T dy. \tag{13}$$

Demostración. La primera afirmación del lema 1 no está relacionada con la dimensión. La segunda se deduce del lema de Fatou y de las relaciones

$$\liminf \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \int \liminf \frac{(\sqrt{f_{\theta'}} - \sqrt{f_{\theta''}})^2}{|\theta' - \theta''|^2} \mu(dx) =$$

$$= \frac{1}{4} \int \varphi((x, \theta), \omega)^2 \mu(dx) = \frac{1}{4} \omega I(\theta)\omega^T.$$

Para demostrar (13) indicaremos que

$$\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}} = \frac{1}{2} \int_0^a (\varphi(x, \theta_1 + y\omega), \omega)dy =$$

$$= \frac{a}{2} \int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega)dy;$$

$$r(\theta_1, \theta_2) = \frac{a^2}{4} \int_{\mathscr{X}} \left[\int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega dy \right]^2 \mu(dx) \leqslant$$

$$\leqslant \frac{a^2}{4} \int_{\mathscr{X}} \int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega)^2 dy \mu(dx) =$$

$$= \frac{a^2}{4} \int_0^1 \int_{\mathscr{X}} (\varphi(x, \theta_1 + ay\omega), \omega)^2 \mu(dx)dy = \frac{a^2}{4} \int_0^1 \omega I(\theta_1 + ay\omega)\omega^T dy. \quad \triangleleft$$

Pongamos, como antes, $r(\Delta) = r(\theta, \theta + \Delta)$. Del lema 1A se deduce el

Teorema 2A. *Si la función* $\sqrt{f_\theta(x)}$ *es derivable cuando c.t.* $|\mu|$ *valores de* x, *y la matriz* $I(\theta)$ *es continua, entonces para cualquier vector* ω *de longitud unitaria existe*

$$\lim_{\delta \to \theta} \frac{r(\delta\omega)}{\delta^2} = \frac{1}{4} \omega I(\theta)\omega^T.$$

Al igual que en el caso unidimensional, del lema 1*A* también podemos obtener el corolario siguiente. Designemos por Sp $I(\theta)$ la traza de la matriz $I(\theta)$.

Teorema 3A. *Si se cumplen las condiciones* (A_0), (A_c), *y la matriz* $I(\theta)$ *es positivamente definida en* Θ, $4h \equiv \sup \operatorname{Sp} I(\theta) < \infty$, *entonces existe una constante* $g > 0$ *tal que para todos* $\theta_1, \theta_2 \in \Theta$

$$g \leqslant \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant h. \tag{14}$$

Demostración. Designemos por $\Lambda_1(\theta)$ y $\Lambda_k(\theta)$ los números propios, mínimo y máximo, respectivamente, de la matriz $I(\theta)$, así que cuando $|\omega| = 1$,

$$\Lambda_1(\theta) \leqslant \omega I(\theta)\omega^T \leqslant \Lambda_k(\theta). \tag{15}$$

Según las condiciones del teorema, $\Lambda_1(\theta) > 0$ siempre en Θ. Como $(\varphi, \omega)^2 \leqslant |\varphi|^2 = \sum_{j=1}^{k} \varphi_j^2$, entonces

$$\int_{\mathcal{X}} (\varphi, \omega)^2 \mu(dx) = \omega I(\theta)\omega^T \leqslant \operatorname{Sp} I(\theta)$$

y, por consiguiente, $\Lambda_k(\theta) \leqslant \operatorname{Sp} I(\theta) \leqslant 4h$. De la desigualdad (13) obtenemos

$$\frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant \frac{1}{4} \int_0^1 \Lambda_k(\theta_1 + ay\omega)dy \leqslant h.$$

Demostremos ahora la segunda desigualdad en (14). Supongamos que ésta no es cierta. Entonces, al igual que en el teorema 3, habrá una sucesión $(\theta_1^{(n)}, \theta_2^{(n)})$, $\theta_1^{(n)} \to \theta_1 \in \Theta$, $\theta_2^{(n)} \to \theta_2 \in \Theta$, para la cual será válida (11). Si $\theta_1 \neq \theta_2$, esto contradirá (5). Si $\theta_1 = \theta_2 = \theta$, entonces, en virtud de la compacticidad de la esfera $|\omega| = 1$, se puede considerar, sin limitar la generalidad, que $\theta_2^{(n)} = \theta_1^{(n)} + \delta\omega^{(n)}$, $\omega^{(n)} \to \omega$, $|\omega^{(n)}| = |\omega| = 1$. Pero en este caso (11) contradirá (12) y (15). ◁

5.* Relación entre las distancias sujetas a examen y las estimaciones. Examinemos la distancia de Kullback—Leibler entre la distribución $\mathbf{P}_\theta$ y la distribución $\mathbf{G}$ que no depende de θ:

$$\varrho_1(\mathbf{G}, \mathbf{P}_\theta) = \int \ln \frac{d\mathbf{G}}{d\mu}\, \mathbf{G}(dx) - \int \ln f_\theta(x)\mathbf{G}(dx).$$

Aquí sólo depende de θ el segundo sumando

$$d(\mathbf{P}_\theta, \mathbf{G}) = -\int \ln f_\theta(x)\mathbf{G}(dx).$$

Por otro lado, recordemos que la e.v.m. ha sido definida en el § 6 como valor de θ con el que se minimiza $d(\mathbf{P}_\theta, \mathbf{P}_n^*)$. Si la distribución de x_1 es discreta, y μ es la medida de cálculo, entonces la expresión

$$d(\mathbf{P}_n^*, \mathbf{P}_n^*) = -\int \ln \frac{d\mathbf{P}_n^*}{d\mu}\, \mathbf{P}_n^*(dx)$$

tiene sentido, $\varrho_1(\mathbf{P}_n^*, \mathbf{P}_\theta) = d(\mathbf{P}_\theta, \mathbf{P}_n^*) - d(\mathbf{P}_n^*, \mathbf{P}_n^*)$ y, por consiguiente, podemos considerar que la e.v.m. minimiza la distancia de Kullback—Leibler

$\varrho_1(\mathbf{P}_n^*, \mathbf{P}_\theta)$ entre $\mathbf{P}_\theta$ y $\mathbf{P}_n^*$. En el caso general tal interpretación puede ser aceptada sólo convencionalmente.

Para las distribuciones discretas de x_1 también se pueden examinar las distancias $\varrho_i(\mathbf{P}_\theta, \mathbf{P}_n^*)$ cuando $i = 2, 3$, así como las estimaciones que minimizan estas distancias. Por ejemplo, cuando $i = 2$ obtenemos

$$\varrho_2(\mathbf{P}_\theta, \mathbf{P}_n^*) = \sum_i \frac{\left(\frac{\nu_i}{n} - f_\theta(a_i)\right)^2}{f_\theta(a_i)},$$

donde ν_i es el número de elementos de la muestra, los cuales han caído en el punto a_i, para el cual $f_\theta(a_i) = \mathbf{P}_\theta\{a_i\} > 0$. Esta es la estadística χ^2 (véase los §§ 7 y 8), debido a lo cual también hemos dado tal denominación a la distancia ϱ_2.

En vista de que las distancias ϱ_i poseen propiedades asintóticas semejantes, las estimaciones que las minimizan, como será aclarado más tarde, coincidirán asintóticamente.

§ 22.* Desigualdad de diferencias del tipo Rao—Cramer

Este párrafo está un poco apartado de la exposición principal. Aquí trataremos de responder, aunque sea parcialmente, a la pregunta acerca de qué es lo que ocurre con la frontera inferior admisible para $\mathbf{M}_\theta(\theta^* - \theta)^2$ en el caso irregular, o sea, en el caso cuando la función $f_\theta(x)$ no es derivable respecto a θ o cuando $I(\theta) = \infty$.

Comenzaremos por el ejemplo que muestra que, en estas condiciones, el comportamiento de las desviaciones estándar (o de sus varianzas) puede diferenciarse totalmente del segundo miembro de la desigualdad de Rao—Cramer.

Ejemplo 1. Sea $X \in \mathbf{U}_{0\theta}$. Aquí, la condición (R) no se cumple, ya que la función $f_\theta(x)$ es discontinua. Como sabemos, para esta familia estadística $S = \max x_i$ es completa y suficiente (véase el ejemplo 14.3). Tomemos la estimación no desplazada $\theta^* = 2x_1$. Entonces, en virtud de los resultados obtenidos en el § 14, la estadística $\theta_S^* = 2\mathbf{M}_\theta(x_1/S)$ será eficiente. Calculemos el valor de $\mathbf{M}_\theta(x_1/S)$. Como $\mathbf{P}_\theta(S < z) = (z/\theta)^n$, $z \in [0, \theta]$, entonces S tiene una densidad igual a nz^{n-1}/θ^n en $[0, \theta]$ e igual a cero fuera de ese intervalo. Para hallar la distribución condicional $P(B/s) = \mathbf{P}_\theta(x_1 \in B/S) = s)$ de la magnitud x_1, a condición de que $S = s$, utilizaremos la regla (10.2):

$$P(dy/s) = \mathbf{P}_\theta(x_1 \in dy/S = s) = \frac{\mathbf{P}_\theta(x_1 \in dy,\ S \in ds)}{\mathbf{P}_\theta(S \in dx)}.$$

Aquí el numerador es igual a

$$\mathbf{P}_\theta(x_1 \in dy,\ S \in dx) = \begin{cases} \dfrac{dy}{\theta} \cdot \dfrac{(n-1)s^{n-2}ds}{\theta^{n-1}} & \text{cuando } y < s, \\ \dfrac{ds}{\theta} \cdot \dfrac{s^{n-1}}{\theta^{n-1}} & \text{cuando } y = s, \\ 0 & \text{cuando } y > s. \end{cases}$$

De aquí se deduce que $P(dy/s) = \dfrac{(n-1)dy}{ns}$ cuando $0 \leqslant y < s$, $P(\{s\}/s) = 1/n$. Por lo tanto,

$$\mathbf{M}_\theta x_1/S) = \int_0^\theta y \frac{n-1}{nS}\,dy + \frac{S}{n} = \frac{S(n-1)}{2n} + \frac{S}{n} = \frac{n+1}{2n}\,S,$$

$$\theta_S^* = S\left(1 + \frac{1}{n}\right).$$

Tenemos

$$\mathbf{D}_\theta\theta_S^* = \mathbf{M}_\theta(\theta_S^*)^2 - \theta^2 = \int_0^\theta s^2\left(1+\frac{1}{n}\right)^2 \frac{ns^{n-1}}{\theta^n}\,ds - \theta^2 =$$

$$= \left(\frac{(n+1)^2}{n(n+2)} - 1\right)\theta^2 = \frac{\theta^2}{n(n+2)}. \quad (1)$$

Como θ_S^* es eficiente, para toda estimación no desplazada θ^*,

$$\mathbf{D}_\theta\theta^* \geqslant \frac{\theta^2}{n(n+2)}. \quad (2)$$

Ahora bien, para grandes valores de n, la desviación estándar de $\mathbf{M}_\theta(\theta_S^* - \theta)^2$ tendrá un orden de pequeñez de $1/n^2$. Desde el punto de vista de la frontera inferior de la desigualdad de Rao—Cramer, que tiene un orden de $1/n$, la misma constituye una exactitud anormalmente alta*). Se puede mostrar que ésta es la exactitud con la que, a partir de la muestra, se determinan cualesquiera puntos de saltos de $f_\theta(x)$ prohibidos por la condición (R)). En el ejemplo 7.4, dedicado a la estimación de la mediana, hemos visto que los puntos donde la densidad $f_\theta(x)$ es infinita, se pueden determinar aún más exactamente, así que, en términos generales, cuanto mayor sea la alteración de la regularidad en el punto, tanto más exactamente será apreciado este punto por la muestra. Digamos, si $X \in \mathbf{P}_\theta$, donde $\mathbf{P}_\theta = \frac{1}{2}\,\mathbf{U}_{0,\theta} + \frac{1}{2}\,\mathbf{I}_\theta$, $\mathbf{I}_\theta$ es la distribución concentrada en el punto θ, entonces

*) Para el parámetro θ también existen estimaciones cuya varianza tiene el orden de $1/n$. Por ejemplo, para la estimación $\theta^{**} = 2\bar{x}$ tenemos $\mathbf{M}\theta^{**} = \theta$, $\mathbf{D}\theta^{**} = \dfrac{4}{n}\,Dx_1 = \dfrac{\theta^2}{3n}$.

$\mathbf{P}_\theta(S \neq \theta) = 2^{-n}(S = \max x_i)$, así que la varianza de $\theta^* - \theta$, cuando $\theta^* = S$, decrecerá exponencialmente con el aumento de n.

¿Será posible en estas condiciones indicar la frontera inferior para la varianza de las estimaciones? Más adelante obtendremos una desigualdad análoga a la de Rao—Cramer, mediante la cual tales fronteras pueden ser construidas cuando las condiciones de regularidad son menos rigurosas que la condición (R).

Solamente supondremos que se cumple la condición (A_μ), aunque tampoco eso tiene mucha importancia (véase la observación al final del párrafo).

Designemos por $\Delta\varphi(\theta)$ el incremento de la función $\varphi(\theta)$ en el intervalo $(\theta, \theta + \Delta)$; por $N^n_{P_\theta}$, el portador en $\mathscr{X}^n$ de la distribución de la muestra: $N^n_{P_\theta} = \{x: f_\theta(x) \neq 0\}$ y pongamos $N^n = N^n_{P_\theta} \cup N^n_{P_{\theta+\Delta}}$.

Teorema 1. (Desigualdad de Chapman—Robbins). *Sea* $\theta \in \Theta$, $\theta + \Delta \in \Theta$, $a(\theta) = \mathbf{M}_\theta\theta^*$ *Entonces, para cualquier* $\Delta \neq 0$,

$$\mathbf{D}_\theta\theta^* \geqslant \frac{(\Delta_a(\theta))^2}{\int[\Delta f_\theta(x)]^2/f_\theta(x)\mu^n(dx)} = \frac{(\Delta a(\theta))^2}{\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta)}, \tag{3}$$

donde ϱ_2 *es la distancia* χ^2 *examinada en el § 21. Aquí, para las estimaciones no desplazadas es necesario sustituir el numerador por* Δ^2.

En virtud del teorema 21.1, el denominador en (3) tiene la forma $\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta) = (1 + r_2(\Delta))^n - 1$, donde

$$r_2(\Delta) = \varrho_2(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta) = \int \frac{[\Delta f_\theta(x)]^2}{f_\theta(x)}\,\mu(dx). \tag{4}$$

Ahora bien, cuanto mayor sea la distancia $\varrho_2(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta)$ entre $\mathbf{P}_{\theta+\Delta}$ y $\mathbf{P}_\theta$ (al ser registrado Δ), tanto menor será la frontera inferior para $\mathbf{D}\theta^*$.

Si $\mathbf{P}_{\theta+\Delta}$ es absolutamente continua respecto a $\mathbf{P}_\theta$ entonces $N^n_{P_{\theta+\Delta}} \subset N^n_{P_\theta} = N^n$ $\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta)$ puede escribirse en la forma (véase (21.2))

$$\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta) = \mathbf{M}_\theta\left[\frac{\Delta f_\theta(X)}{f_\theta(X)}\right]^2;$$

análogamente, $r_2(\Delta) = \mathbf{M}_\theta\left[\dfrac{\Delta f_\theta(x_1)}{f_\theta(x_1)}\right]^2$.

Pero si la distribución $\mathbf{P}_{\theta+\Delta}$ no es absolutamente continua respecto a $\mathbf{P}_\theta$, entonces existe un subconjunto de $N_{P_{\theta+\Delta}}$ de medida positiva $\mathbf{P}_{\theta+\Delta}$ en el que $f_\theta(x) = 0$, así que la integral en (4) se vuelve infinita, y la propia desigualdad (3) se vuelve trivial. Es necesario señalar otra vez, que en este caso la expresión $\mathbf{M}_\theta[\Delta f_\theta(X)/f_\theta(X)]^2$, entendida como integral respecto a N_{P_θ}, puede permanecer finita.

Demostración del teorema 1. De lo dicho anteriormente se deduce que, sin limitar la generalidad, podemos considerar que $\mathbf{P}_{\theta+\Delta}$ es absolutamente continua respecto a $\mathbf{P}_\theta$, así que $N^n_{P_{\theta+\Delta}} \subset N^n_{P_\theta} = N^n$. Como $f_\theta(x)$ y $f_{\theta+\Delta}(x)$ es la densidad en $\mathscr{X}^n$, entonces

$$\int \Delta f_\theta(x)\mu^n(dx) = 0.$$

Además,

$$\int \theta^* \Delta f_\theta(x)\mu^n(dx) = \Delta a(\theta).$$

De aquí se desprende que

$$\int_{N^n} (\theta^* - a(\theta))\Delta f_\theta(x)\mu^n(dx) = \Delta a(\theta). \tag{5}$$

En el conjunto N^n podemos representar la función subintegral de (5) en forma del producto

$$(\theta^* - a(\theta))^2\sqrt{f_\theta(x)} \cdot \frac{\Delta f_\theta(x)}{\sqrt{f_\theta(x)}}.$$

Aplicando luego la desigualdad de Cauchy—Buniakovski, obtenemos

$$(\Delta a(\theta))^2 \leqslant \int_{N^n} (\theta^* - \theta)^2 f_\theta(x)\mu^n(dx) \int_{N^n} \frac{(\Delta f_\theta(x))^2}{f_\theta(x)}\,\mu^n(dx). \quad \triangleleft$$

En lo sucesivo, según las observaciones hechas más arriba, nos limitaremos, al igual que en la demostración del teorema 1, al caso cuando $\mathbf{P}_{\theta+\Delta}$ es absolutamente continua respecto a $\mathbf{P}_\theta$ (de lo contrario la desigualdad (3) se vuelve trivial).

Corolario 1. *Si se cumplen las condiciones de regularidad que aseguran la existencia* (véase la observación 21.1 al teorema 21.2) de $\lim_{\Delta\to 0} r_2(\Delta)/\Delta^2 = I(\theta)$, *entonces*

$$\mathbf{D}_\theta\theta^* \geqslant \frac{(a'_+(\theta))^2}{nI(\theta)}, \tag{6}$$

donde $a'_+(\theta) = \lim_{\Delta\to 0} \sup \frac{\Delta a(\theta)}{\Delta}$.

Para obtener (6) del teorema 1 sólo es necesario notar que podemos elegir la sucesión $\Delta \to 0$ de modo que $\frac{\Delta a(\theta)}{\Delta} \to a'_+(\theta)$. $\triangleleft$

La desigualdad (6) es, según su forma, cierta generalización de la desigualdad de Rao—Cramer (generalización, lo más probable, ficticia, ya que las condiciones de regularidad mencionadas conducen, por lo visto, a la existencia de $a'(\theta)$).

La desigualdad (3), por supuesto, se denomina desigualdad de *diferencias*, a distinción de la desigualdad (6) que podría denominarse desigualdad *diferencial*.

Ahora bien, si $r_2(\Delta) \sim I(\theta)\Delta^2$ (esto corresponde al hecho de que f_θ es derivable), *entonces de la desigualdad de diferencias de Chapman—Robbins se deduce la desigualdad diferencial de Rao—Cramer.*

Pero si la función f_θ no es derivable, entonces, al disminuir Δ, el comportamiento de $r_2(\Delta)$ será diferente.

Si, digamos, f_θ es derivable en todas partes, a excepción de un número finito de puntos de discontinuidad $\theta = \theta(x)$ que dependen de x, entonces tendremos

$$r_2(\Delta) \sim c|\Delta|. \tag{7}$$

Esto puede ser aclarado de la forma más sencilla a base de un ejemplo muy típico, examinado al principio del párrafo.

Sea $X \in \mathbf{U}_{0,\theta}$. Para que sea cumplida la condición de continuidad absoluta de $\mathbf{P}_{\theta+\Delta}$ respecto a $\mathbf{P}_\theta$, en el caso de $\mathbf{P}_\theta = \mathbf{U}_{0,\theta}$ consideraremos que $\Delta < 0$, $|\Delta| < \theta$. Entonces

$$\Delta f_\theta(x) = \begin{cases} \dfrac{1}{\theta+\Delta} - \dfrac{1}{\theta} & \text{para } x \in [0,\ \theta+\Delta], \\ -\dfrac{1}{\theta} & \text{para } x \in [0+\Delta,\ \theta], \\ 0 & \text{para } x \notin [0,\ \theta], \end{cases}$$

$$r_2(\Delta) = \int_0^\theta \frac{(\Delta f_\theta(x))^2}{f_\theta(x)}\,dx = \int_0^{\theta+\Delta} \left[\frac{\Delta}{\theta(\theta+\Delta)}\right]^2 \theta\,dx = \int_{\theta+\Delta}^{\theta} \frac{1}{\theta^2}\,\theta\,dx =$$

$$= \frac{\Delta^2}{\theta(\theta+\Delta)} + \frac{|\Delta|}{\theta}.$$

Lo esencial aquí es la existencia del intervalo cuya longitud es comparable con Δ y en el que $|\Delta f_\theta(x)| > c > 0$, donde c no depende de Δ. Esto asegura precisamente el orden de pequeñez (7) para $r_2(\Delta)$.

Volviendo a nuestro ejemplo, vemos que para las estimaciones no desplazadas del parámetro θ,

$$\mathbf{D}\theta^* \geqslant \max_\Delta \frac{\Delta^2}{\left(1 + \dfrac{|\Delta|}{\theta} + \dfrac{\Delta^2}{\theta(\theta+\Delta)}\right)^n - 1}.$$

¿Cuál es el orden de pequeñez del segundo miembro de esta desigualdad cuando $n \to \infty$? Suponiendo $|\Delta| = y\theta/n$, obtenemos

$$\mathbf{D}\theta^* \geqslant \frac{\theta^2}{n^2} \max_y \frac{y^2}{\left(1 + \frac{y}{n} + \frac{y^2}{n(n-y)}\right)^n - 1}.$$

Está claro que la expresión con signo máx es asintóticamente equivalente a $h = \max y^2/(e^y - 1) \approx 0{,}65$, así que

$$\mathbf{D}\theta^* \geqslant \frac{\theta^2}{n^2}(h + o(1)).$$

En cuanto al orden de pequeñez, esta desigualdad tiene el mismo segundo miembro que la desigualdad inmejorable (2), pero el factor constante de θ^2/n^2 en (2) es "mejor" y es igual a 1.

A la par con (7) pueden aparecer también otras velocidades de convergencia de $r_2(\Delta)$ hacia el cero, cuando $\Delta \to 0$. Podemos obtener, por ejemplo, tanto $r_2(\Delta) \sim c\Delta^\alpha$, $\alpha < 1$, si $f_\theta(x)$ tiene líneas de $\theta = \theta(x) \neq \text{const}$, al aproximarse a las cuales $f_\theta(x) \to \infty$; como también $r_2(\Delta) \sim c\Delta^\alpha$, $2 > \alpha > 1$, si f_θ es continua respecto a θ pero no es derivable sino satisface solamente la condición de Hölder en el entorno de cierta línea $\theta = \theta(x) \neq \text{const}$. No es difícil ver que el orden de pequeñez

$$\max_\Delta \frac{\Delta^2}{(1 + c\Delta^\alpha)^n - 1}$$

para $\alpha < 2$ será definido por el valor de $\Delta = (y/cn)^{1/\alpha}$, así que

$$\mathbf{D}\theta^* \geqslant \frac{1}{(cn)^{2/\alpha}} \max_y \frac{y^{2/\alpha}}{e^y - 1}(1 + o(1)).$$

En el caso "regular" $\alpha = 2$, el máximo respecto a y se obtiene en el punto límite $y = 0$ ($\Delta = 0$).

Concluyendo este párrafo señalaremos que las estimaciones para $\mathbf{D}\theta^*$ también pueden ser obtenidas, de modo análogo, para las no absolutamente bicontinuas $\mathbf{P}_\theta$ y $\mathbf{P}_{\theta+\Delta}$. Para esto, en (5) es necesario multiplicar y dividir la función subintegral no por $\sqrt{f_\theta(x)}$, sino por $\sqrt{f_\theta(x) + f_{\theta+\Delta}(x)}$. La condición (A_μ) tampoco es tan esencial, ya que las medidas de $\mathbf{P}_{\theta_i}$ y $\mathbf{P}_{\theta+\Delta}$ siempre son absolutamente continuas respecto a $\frac{1}{2}(\mathbf{P}_\theta + \mathbf{P}_{\theta+\Delta})$.

§ 23. Desigualdades auxiliares para la relación de verosimilitud. Conciliabilidad de las estimaciones de la verosimilitud máxima

En los §§ 12—16 hemos estudiado las cuestiones relacionadas con la existencia y la determinación, en forma explícita, de las estimaciones eficientes y R-eficientes. Hemos visto que éstas existen no siempre, ni mucho menos, y pueden ser halladas tan sólo en el caso cuando la función de verosimilitud tiene una forma especial o cuando conocemos, de manera explícita, la estadística suficiente completa (la primera de estas condiciones a menudo conduce a la segunda (véase el § 15)).

Pasemos ahora a la construcción de las estimaciones asintóticamente óptimas. Aquí las condiciones de su existencia serán mucho más amplias. Los resultados respectivos se apoyan, ante todo, en las propiedades asintóticas de la función

$$Z(u) = \frac{f_{\theta+u}(X)}{f_\theta(X)} = \exp\{L(X, \theta + u) - L(X, \theta)\}, \qquad (1)$$

donde, como antes, $L(X, \theta) = \sum_{i=1}^{n} l(x_i, \theta)$. Por regla general, el número θ en (1) se considerará registrado y representará el valor real del parámetro, o sea, tal que $X \in \mathbf{P}_\theta$. En este caso $Z(u)$ es la función de los variables u y X y, por lo tanto, junto con la función de verosimilitud $f_{\theta+u}(X)$, será la *función aleatoria* de la variable u. Llamaremos *relación de verosimilitud* la función $Z(u)$ que desempeña un papel muy importante en la estadística matemática. La tarea principal de este párrafo y del párrafo siguiente consiste en estudiar las propiedades de $Z(u)$.

Será establecido que $Z(u)$ es próxima a cero fuera del entorno del punto $u = 0$. En el entorno de este punto, $Z(u)$ se aproxima, desde cierto punto de vista, a la función delta, mejor dicho, $Z(v/\sqrt{n})$ se aproxima asintóticamente, cuando $n \to \infty$, a la función de densidad de la ley normal.

En los §§ 23—26 examinaremos sólo el parámetro unidimensional. El caso del parámetro multidimensional será investigado separadamente en el § 28.

En las estimaciones posteriores desempeñará un gran papel la distancia de Hellinger

$$r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta) = \int \left(\sqrt{f_{\theta+u}(x)} - \sqrt{f_\theta(x)}\right)^2 \mu(dx)$$

entre las distribuciones $\mathbf{P}_{\theta+u}$ y $\mathbf{P}_\theta$. Hemos examinado esta distancia en el § 21. Recordemos que

$$0 \leqslant r(u) = 2\left(1 - \int \sqrt{f_{\theta+u}(x) f_\theta(x)}\, \mu(dx)\right) \leqslant 2,$$

así que

$$\mathbf{M}_\theta \sqrt{\frac{f_{\theta+u}(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)}} = \int \sqrt{f_{\theta+u}(x) f_\theta(x)}\, \mu(dx) = 1 - r(u)/2, \tag{2}$$

$$\mathbf{M}_\theta Z^{1/2}(u) = (1 - r(u)/2)^n. \tag{3}$$

En lo que se refiere a la familia paramétrica $\{\mathbf{P}_\theta\}$, supondremos en este párrafo y en los párrafos siguientes que *a la par con* (A_μ) *se cumplen las condiciones* (A_0) $(f_{\theta_1}(x) \neq f_{\theta_2}(x)$ *para* $\theta_1 \neq \theta_2$ y (A_c) $(\Theta$ *es un compacto*). El hecho de que la última condición es poco importante desde el punto de vista de aplicaciones, ha sido mencionado anteriormente. Esto se debe a que en los problemas reales, de ordinario es posible señalar las fronteras de los posibles valores de θ, partiendo de las consideraciones a priori. Para simplificar la exposición, allí donde sea necesario, también supondremos que Θ es convexo (en el caso unidimensional esto quiere decir que $\Theta = [a, b]$, $-\infty < a < b < \infty$).

Además, en este párrafo supondremos que la función $\sqrt{f_\theta}$ es *derivable para c.t.* $[\mu]$ *valores de x, y que la información de Fisher*

$$I(\theta) = \int \frac{f'_\theta(x))^2}{f_\theta(x)}\, \mu(dx) = \mathbf{M}_\theta \left(\frac{f'_\theta(\mathbf{x}_1)}{f_\theta(\mathbf{x}_1)}\right)^2$$

es estrictamente positiva y está limitada en Θ. En estas condiciones hemos demostrado en el teorema 21.3 que para todos θ y $\theta + u$ admisibles (o sea, tales que $\theta \in \Theta$, $\theta + u \in \Theta$) para la magnitud $r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta)$ es válida la desigualdad

$$\inf_{\theta, u} \frac{r(u)}{u^2} \geqslant g > 0. \tag{4}$$

1. Desigualdades principales. Designemos, para abreviar, $p(u) = Z^{3/4}(u)$ y supongamos que se cumplen todas las condiciones anteriormente citadas.

Teorema 1.

$$\mathbf{M}_\theta Z^{1/2}(u) \leqslant e^{-ngu^2/2}, \quad \mathbf{M}_\theta p(u) \leqslant e^{-ngu^2/4}, \tag{5}$$

$$\mathbf{M}_\theta |p'(u)| \leqslant \frac{3}{4} \sqrt{nI(\theta + u)}\, e^{-u^2 ng/4}.$$

De las investigaciones realizadas en el § 21 se deduce que para los valores $u = o(1)$ en estas desigualdades, en vez de g se pueden tomar los valores tan próximos como se quiera a $I(\theta)$.

Demostración. En virtud de (3) y (4) tenemos

$$\mathbf{M}_\theta Z^{1/2}(u) = (1 - r(u)/2)^n \leqslant \exp\{-nr(u)/2\} \leqslant \exp\{-ngu^2/2\}.$$

Luego, en virtud de la desigualdad de Cauchy — Buniakovski,

$$\mathbf{M}_\theta p(u) \leqslant [\mathbf{M}_\theta Z^{1/2}(u) \cdot \mathbf{M}_\theta Z(u)]^{1/2} = [\mathbf{M}_\theta Z^{1/2}(u)]^{1/2} \leqslant e^{-u^2 ng/4}.$$

Volviendo a utilizar la desigualdad de Cauchy — Buniakovski y la relación

$$p'(u) = \frac{3}{4} L'(X, \theta + u) Z^{3/4}(u),$$

hallamos

$$\mathbf{M}_\theta |p'(u)| = \frac{3}{4} \mathbf{M}_\theta |L'(X, \theta +)| Z^{1/2}(u) Z^{1/4}(u) \leqslant$$

$$\leqslant \frac{3}{4} [\mathbf{M}_\theta [L'(X, \theta + u)]^2 Z(u) \cdot \mathbf{M}_\theta Z^{1/2}(u)]^{1/2} \leqslant$$

$$\leqslant \frac{3}{4} [\mathbf{M}_{\theta+u} [L'(X, \theta + u)]^2]^{1/2} e^{-u^2 ng/4}. \quad \triangleleft$$

Teorema 2. *Para todos* z, $n \geqslant 1$

$$\mathbf{P}_\theta \Big(\sup_{|v| \geqslant u} Z(v/\sqrt{n}) \Big) \leqslant c e^{-3z/4} e^{-u^2 g/4},$$

donde $c = 2 + 3\sqrt{\pi I_0 / g}$, $I_0 = \sup_{\theta \in \Theta} I(\theta)$ *no dependen de* θ.

Para demostrar el teorema necesitaremos el

Lema 1. *Para todos* $x \geqslant 0$,

$$\int_x^\infty e^{-v^2/2} dv \leqslant \sqrt{2\pi}\, e^{-x^2/2}.$$

Demostración*). La función característica de la variable aleatoria $\xi \in \Phi_{0,1}$ es igual a $\mathbf{M}e^{it\xi} = e^{-t^2/2}$ y está definida en todo el plano. Suponiendo $t = -ix$, obtendremos $\mathbf{M}e^{x\xi} = e^{x^2/2}$. De aquí, con ayuda de la desigualdad de Chébishev, obtenemos

$$\mathbf{P}(\xi > x) = \mathbf{P}(e^{\xi x} > e^{x^2}) \leqslant e^{-x^2} \mathbf{M} e^{\xi x} = e^{-x^2/2}. \quad \triangleleft$$

Demostración del teorema 2. Estimemos la función

$$H(\delta) = \mathbf{M}_\theta \sup_{|v| > \delta} p(v).$$

Si $v \in [\theta + \delta, b]$, entonces

$$p(v - \theta) = p(\delta) + \int_\delta^{v-\theta} p'(u) du \leqslant p(\delta) + \int_\delta^{b-\theta} |p'(u)| du.$$

---

*) Para grandes x son más exactas las desigualdades siguientes;

$$\frac{1}{x+1} e^{-x^2/2} < \int_x^\infty e^{-v^2/2} dv < \frac{1}{x} e^{-x^2/2},$$

las cuales pueden ser fácilmente obtenidas por el lector, comparando las derivadas de las funciones sujetas a examen (los valores de las propias funciones coinciden cuando $x = \infty$).

Como aquí el segundo miembro no depende de v, entonces

$$\sup_{u \geqslant \delta} p(u) \leqslant p(\delta) + \int\limits_{u \geqslant \delta} |p'(u)|du,$$

$$H_+(\delta) \equiv \mathbf{M}_\theta \sup_{u \geqslant \delta} p(u) \leqslant \mathbf{M}_\theta p(\delta) + \int\limits_{u \geqslant \delta} \mathbf{M}_\theta |p'(u)|du.$$

De aquí, en virtud del teorema 1 obtenemos

$$H_+(\delta) \leqslant e^{-\delta^2 ng/4} + \frac{3}{4}\sqrt{n} \int\limits_{u \geqslant \delta} \sqrt{I(\theta + u)}\, e^{-u^2 ng/4} du.$$

A base del lema 1,

$$H_+(\delta) \leqslant e^{-ng\delta/4} + \frac{3}{4}\sqrt{nI_0} \int\limits_{|u| \geqslant \delta} e^{-ngu^2/4} du \leqslant$$

$$\leqslant e^{-ng\delta^2/4} + \frac{3}{4}\sqrt{2I_0/g} \int\limits_{v \geqslant \delta\sqrt{ng/2}} e^{-v^2/2} dv \leqslant e^{-ng\delta^2/4}\left(1 + \frac{3}{2}\sqrt{\pi I_0/g}\right).$$

Está claro que una estimación exactamente igual, será válida para la función

$$H_-(\delta) = \sup_{u \leqslant -\delta} p(u).$$

Por eso

$$H(\delta) \leqslant H_+(\delta) + H_-(\delta) \leqslant (2 + 3\sqrt{\pi I_0/g})e^{-ng\delta^2/4}.$$

Queda hacer uso de la desigualdad de Chébishev:

$$\mathbf{P}_\theta(\sup_{|t| \geqslant \delta} Z(t) > e^z) = \mathbf{P}_\theta(\sup_{|t| \geqslant \delta} p(t) > e^{3z/4}) \leqslant H(\delta)e^{-3z/4}. \quad \triangleleft$$

2. Estimaciones para la distribución y los momentos de la e.v.m. Conciliabilidad de la e.v.m.

Teorema 3. *Existen valores de* $c < \infty$, $g > 0$ *tales, que*

$$\mathbf{P}_\theta(\sqrt{n}(\hat{\theta}^* - \theta) \geqslant v) \leqslant ce^{-gv^2/4} \tag{6}$$

para todos v *y* $n \geqslant 1$.

Demostración. Del teorema 2 se desprende que

$$\mathbf{P}_\theta(\sup_{|t| \geqslant v/\sqrt{n}} Z(t) > 1) \leqslant ce^{-gv^2/4}.$$

Queda hacer uso de la relación

$$\{|\hat{\theta}^* - \theta| \geqslant \delta\} = \left\{\sup_{|t| \geqslant \delta} Z(t) \geqslant \sup_{|t| \leqslant \delta} Z(t)\right\} \subset \left\{\sup_{|t| \geqslant \delta} Z(t) \geqslant Z(0) = 1\right\} \tag{7}$$

cuando $\delta = v\sqrt{n}$. $\triangleleft$

Corolario 1. *Supongamos que* $u_n \to \infty$ *es toda sucesión indefinidamente creciente. Entonces*

$$(\hat{\theta}^* - \theta)\sqrt{n}/u_n \underset{P}{\to} 0. \tag{8}$$

No obstante, si u_n *son tales que para cualquier* $\alpha > 0$

$$\sum e^{-\alpha u_n^2} < \infty, \tag{9}$$

entonces

$$(\hat{\theta}^* - \theta)\sqrt{n}/u_n \underset{\text{c.s.}}{\to} 0. \tag{10}$$

Estas relaciones son, evidentemente, las amplificaciones de la conciliabilidad ($\hat{\theta}^* - \theta \underset{P}{\to} 0$) y de la conciliabilidad fuerte ($\hat{\theta}^* - \theta \underset{\text{c.s.}}{\to} 0$) de la e.v.m., respectivamente.

Demostración. La relación (8) se deduce directamente de (6) si en esta última se pone $v = \delta u_n$. La relación (10) también se desprende de (6), ya que la suma de los segundos miembros en (6), al cumplirse (9), formará una serie convergente. ◁

Por ejemplo, incluso una sucesión tan lentamente creciente como $u_n = \ln n$ satisface la condición (9), así que*)

$$(\hat{\theta}^* - \theta)\sqrt{n}/\ln n \underset{\text{c.s.}}{\to} 0.$$

Corolario 2. *Existe un valor* $c_1 < \infty$, *no dependiente de* n *y* θ, *tal que para todo* $\alpha \leqslant g/5$,

$$\mathbf{M}_\theta \exp\{\alpha(u^*)^2\} < c_1, \text{ donde } u^* = \sqrt{n}(\hat{\theta}^* - \theta). \tag{11}$$

Demostración. Integrando por partes, obtenemos

$$\mathbf{M}e^{\alpha\xi^2} = -\int_0^\infty e^{\alpha v^2} d\mathbf{P}(|\xi| \geqslant v) = 1 + 2\alpha \int_0^\infty v e^{\alpha v^2} \mathbf{P}(|\xi| \geqslant v) dv.$$

Por eso, en virtud del teorema 3,

$$\mathbf{M}_\theta e^{\alpha(u^*)^2} \leqslant 1 + \frac{2g}{5} \int_0^\infty v e^{-gv^2/20} dv \equiv c_1 < \infty. \quad \triangleleft$$

§ 24. Propiedades asintóticas de la relación de verosimilitud

En el párrafo precedente hemos establecido una serie de desigualdades para $Z(u)$. Determinemos ahora la distribución límite para tales funciones aleatorias. Esto se hace cuando se cumpla la condición (R) del § 16. No obstante, para simplificar los razonamientos, introduzcamos ciertas

*) De la observación 25.2 resultará que (10) también es válida para u_n que crecen aún más lentamente.

suposiciones adicionales que no siempre están relacionadas con la esencia de la cuestión, pero hacen más breves y más claras las demostraciones.

Designemos con el símbolo (RR), las condiciones introducidas para indicar asimismo que tales son las condiciones de regularidad y que ellas intensifican las condiciones (R).

Condiciones (RR):

1) se *cumplen las condiciones* (A_0), (A_c), (R).

2) *la función* $l(x, \theta)$ *para c.t.* $[\mu]$ *valores de* x *es dos veces continuamente derivable respecto a* θ. *La función* $|l''(x, t)|$ *es mayorada por la función* $l(x)$ *que no depende de* t: $|l''(x, t)| < l(x)$, *para la cual la integral*

$$\mathbf{M}_t l(\mathrm{x}_1) = \int l(x) f_t(x) \mu(dx)$$

converge uniformemente en $t \in \Theta$*)*.

Por convergencia uniforme de la integral entendemos la convergencia**)

$$\sup_{\theta} \int\limits_{x:|l(x)|>N} l(x) f_\theta(x) \mu(dx) \to 0$$

cuando $N \to \infty$.

Posteriormente necesitaremos las dos propiedades siguientes, que se deducen de (RR):

1) *Validez de la derivación doble respecto al parámetro bajo el signo de integral en la igualdad*

$$\int f_\theta(x) \mu(dx) = 1$$

que significa la validez de las relaciones

$$\int f'_\theta(x) \mu(dx) = 0, \quad \int f''_\theta(x) \mu(dx) = 0. \tag{1}$$

2) *Convergencia uniforme de la integral*

$$I(\theta) = \int (l'(x, \theta)^2 f_\theta(x) \mu(dx).$$

(esta propiedad se deduce de (R) y se necesitará en el § 29).

---

*) Toda la exposición ulterior conservará su validez si la condición y la existencia de la mayorante se debilitan del modo siguiente: la región Θ puede ser cubierta por el número finito de regiones $\Theta_1, \ldots, \Theta_s$ de tal modo que cuando $\theta \in \Theta_j$ la función $l''(x, \theta)$ es mayorada por la función $l_{(j)}(x)$ que no depende de t: $|l''(x, \theta)| < l_{(j)}(x)$, para la cual la integral

$$\mathbf{M}_\theta l_{(j)}(\mathrm{x}_1) = \int l_{(j)}(x) f_\theta(x) \mu(dx)$$

converge uniformemente en $\theta \in \Theta_j$, $j = 1, \ldots, s$.

**) Tal comprensión de la convergencia uniforme se halla en concordancia con la convergencia uniforme utilizada en el teorema 1.5.4. Aquí ella pertenecía a la función $l(x) \equiv x$. A su vez, la misma no es la convergencia uniforme $\int \varphi(x, \theta) \mu(dx)$ para $\varphi(x, \theta) = l(x) f_\theta(x)$ cuando se supone que, para $N \to \infty$,

$$\sup_{\theta} \int\limits_{x:|\varphi(x,\theta)|>N} (x, \theta) \mu(dx) \to 0.$$

Para descargar la exposición fundamental, la demostración de estos corolarios de las condiciones (RR) se da en el Suplemento VI. La exposición también se puede simplificar de otra manera: introduciendo en las condiciones (RR) las dos propiedades mencionadas y despreciando el hecho de que en tal forma ellas serán "redundantes".

En vista de que

$$l'(x,\ \theta) = \frac{f'_\theta(x)}{f_\theta(x)}, \quad l''(x,\ \theta) = \frac{f''_\theta(x)}{f_\theta(x)} - \left(\frac{f'_\theta(x)}{f_\theta(x)}\right)^2,$$

la relación (1) se puede escribir en la forma

$$\mathbf{M}_\theta l'(\mathrm{x}_1,\ \theta) = 0,\ \mathbf{M}_\theta l''(\mathrm{x}_1,\ \theta) = -\mathbf{M}_\theta(l'(\mathrm{x}_1,\ \theta))^2 = -I(\theta). \tag{2}$$

Ya hemos utilizado la primera de estas igualdades.

Señalemos un corolario más de las condiciones (RR). Estas últimas son mucho más fuertes que las condiciones utilizadas en los §§ 21 y 23 y, por consiguiente, *tienen lugar todas las afirmaciones de los teoremas del* § 23 *acerca de las estimaciones para la distribución* $\sup_{|v|\geqslant u} Z(v/\sqrt{n})$, *y acerca de la conciliabilidad de la e.v.m.*

Lema 1. *Si se cumplen las condiciones* (RR), *tiene lugar la continuidad* $l''(x,\ \theta)$ *"por término medio" desde el punto de vista siguiente*:

$$\mathbf{M}_\theta \omega''_\Delta(\mathrm{x}_1) = \int \omega''_\Delta(x) f_\theta(x)\mu(dx) \to 0 \tag{3}$$

para $\Delta \to 0$, *donde* $\omega''_\Delta(x)$ *es el módulo de continuidad de la función* $l''(x,\ \theta)$:

$$\omega''_\Delta(x) = \sup_{\substack{\theta\in\Theta,\theta+u\in\Theta \\ \in\Theta \\ |u|\leqslant\Delta}} |l''(x,\ \theta+u) - l''(x,\ \theta)|. \tag{4}$$

Demostración. En virtud del teorema de convergencia mayorable, la relación (3) será el corolario de la continuidad ordinaria, puesto que en este caso $\omega''_\Delta(x) \to 0$ para c.t. $[\mu]$ valores de x cuando $\Delta \to 0$ y, además, $|\omega''_\Delta(x)| \leqslant 2I(x)$. ◁

Designemos

$$\gamma_n(\Delta,\ \theta) = \sup_{|v|\leqslant\Delta} \left| \frac{L'(X,\ \theta+v) - L'(X,\ \theta)}{nv} + I(\theta) \right|.$$

Lema 2. *Supongamos que se cumplen las condiciones* (RR), $\delta_n > 0$, $n = 1, 2, \ldots$, *es cualquier sucesión convergente a cero. Entonces, para cualquier* $\theta \in \Theta$ *y para* $X \in \mathbf{P}_\theta$,

$$\gamma_n(\delta_n,\ \theta) \underset{\text{c.s.}}{\to} 0,\ \gamma_n(\delta_n,\ \hat\theta^*) \underset{\text{c.s.}}{\to} 0.$$

En estas relaciones, $I(\theta)$ *se puede sustituir por* $I(\hat\theta^*)$ *y al contrario.*

Demostración. Demostremos al principio la primera afirmación. Como $\mathbf{M}_\theta l''(\mathrm{x}_1, \theta) = -I(\theta)$, $L''(X, \theta)/n \underset{\text{c.s.}}{\to} -I(\theta)$, es suficiente cerciorarse de que $\gamma_n(\delta_n) \to 0$, donde

$$\gamma_n(\Delta) = \sup_{|v| \leqslant \Delta} \left| \frac{L'(X, \theta + v) - L'(X, \theta)}{nv} - \frac{L''(X, \theta)}{n} \right|.$$

Pero

$$\gamma_n(\delta_n) = \leqslant \sup_{|v| \leqslant \delta_n} \frac{1}{n} |L''(X, \theta + v) - L''(X, \theta)| \leqslant \frac{1}{n} \sum_{i=1}^{n} \omega''_{\delta_n}(\mathrm{x}_i) \equiv \bar{\omega}''_{\delta_n}(X),$$

donde $\omega''_\Delta(x)$ significa el módulo de continuidad $l''(x, \theta)$, definido en (4). Es evidente que para cualquier $\Delta > 0$ registrado, cuando n son bastante grandes,

$$\bar{\omega}''_{\delta_n}(X) \leqslant \bar{\omega}''_\Delta(X).$$

Además, según la ley fuerte de los grandes números,

$$\bar{\omega}''_\Delta(X) \underset{\text{c.s.}}{\to} \mathbf{M}_\theta \omega''_\Delta(\mathrm{x}_1) \equiv \omega''_\Delta.$$

En virtud del lema 1, $\omega''_\Delta \to 0$ cuando $\Delta \to 0$. De aquí se deduce que

$$\bar{\omega}''_{\delta_n}(X) \underset{\text{c.s.}}{\to} 0. \tag{5}$$

La primera afirmación queda demostrada. De (5) y de la definición de la convergencia casi segura se desprende que a la par con (5),

$$\bar{\omega}''_{\delta_n + \eta_n}(X) \underset{\text{c.s.}}{\to} 0$$

para toda sucesión de las variables aleatorias $\eta_n \underset{\text{c.s.}}{\to} 0$. Nos queda señalar que

$$\sup_{|v| < \delta_n} \left| \frac{L'(X, \hat{\theta}^* + v) - L'(X, \hat{\theta}^*)}{nv} - \frac{L''(X, \theta)}{n} \right| \leqslant \bar{\omega}''_{\delta_n + |\hat{\theta}^* - \theta|}(X) \tag{6}$$

y hacer uso del corolario 23.1. La posibilidad de sustituir $I(\theta)$ por $I(\hat{\theta}^*)$ también se deduce del corolario 23.1 (y de la continuidad de $I(\theta)$). ◁

Ahora podemos enunciar las principales afirmaciones acerca del comportamiento asintótico de la relación de verosimilitud $Z(t)$. Designemos

$$Y(u) = \ln Z(u/\sqrt{n}) = L(X, \theta + u/\sqrt{n}) - L(X, \theta)$$

y convengamos en designar por $\varepsilon_n(X, \theta)$ (a veces con índices adicionales) las diferentes sucesiones de variables aleatorias convergentes casi seguramente a cero respecto a $\mathbf{P}_\theta$.

Teorema 1. *Supongamos que se cumplen las condiciones (RR),* $\delta_n > 0$ *es una sucesión arbitraria que converge hacia el cero. Entonces para* $|u/\sqrt{n}| < \delta_n$

$$Y(u) = u\xi_n - \frac{u^2}{2} I(\theta)(1 + \varepsilon_n(X, \theta, u)), \tag{7}$$

donde

$$|\varepsilon_n(X, \theta, u)| \leqslant \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0, \quad \xi_n = L'(X, \theta)/\sqrt{n} \in \Phi_{0, I(\theta)}.$$

El punto $u^* = (\hat{\theta}^* - \theta)\sqrt{n}$, *en el que* $Y(u)$ *alcanza el valor máximo, posee la propiedad*

$$u^* = \frac{\xi_n}{I(\theta)} (1 + \varepsilon_n(X, \theta)), \tag{8}$$

$$2Y(u^*) = 2 \ln Z(\hat{\theta}^* - \theta) = \frac{\xi_n^2}{I(\theta)} (1 + \varepsilon_n(X, \theta)) \in H_1. \tag{9}$$

A la par con (7) *es válida la representación*

$$Y(u) = Y(u^*) - \frac{(u - u^*)^2}{2} I(\theta)(1 + \varepsilon_n(X, \theta, u)), \tag{10}$$

$$|\varepsilon_n(X, \theta, u)| < \varepsilon_n(X, \theta).$$

En todas las afirmaciones dadas se puede sustituir $I(\theta)$ *por* $I(\hat{\theta}^*)$.

En este teorema, al igual que en el lema 2, se supone que $\theta + u\sqrt{n} \in \Theta$. Esta relación será cumplida automáticamente para n bastante grandes si θ es el punto interior de Θ.

Observación. 1. Es importante notar que en (7) las variables aleatorias ξ_n y $\varepsilon_n(X, \theta)$ no dependen de n. Por eso la primera afirmación del teorema puede ser escrita en la forma

$$\sup_{|u| \leqslant \delta_n \sqrt{n}} \left| \frac{Y(u) - u\xi_n + \frac{u^2}{2} I(\theta)}{u^2} \right| \underset{\text{c.s.}}{\to} 0.$$

Si δ_n es tal que

$$\sum_e^{-ng\delta_n^2/4} < \infty, \tag{11}$$

del teorema 23.2 se deduce que en la región adicional $|u| > \delta_n\sqrt{n}$,

$$\sup_{|u| > \delta_n \sqrt{n}} Y(u) \underset{\text{c.s.}}{\to} -\infty.$$

Demostración del teorema 1. Del lema 2 $|v| \leqslant \delta_n$ obtenemos

$$L'(X, \theta + v) = L'(X, \theta) - nvI(\theta)(1 + \varepsilon_n(X, \theta, v)),$$

$$|\varepsilon_n(X, \theta, v)| \leqslant \varepsilon_n(X, \theta).$$

Integrando esta igualdad respecto a v dentro de los límites de 0 a $u/\sqrt{n}$, obtendremos

$$L(X,\ \theta + u/\sqrt{n}) - L(X,\ \theta) = uL'(X,\ \theta)/\sqrt{n} - \frac{u^2}{2} I(\theta)(1 + \varepsilon_n(X,\ \theta,\ u)),$$

$$|\varepsilon_n(X,\ \theta,\ u)| \leqslant \varepsilon_n(X,\ \theta). \tag{12}$$

Esto es, evidentemente, el desarrollo en serie de Taylor, donde $L''(X,\ \theta)/n$ ha sido sustituida por $I(\theta)$, y el término residual admite una estimación uniforme. En vista de que

$$\xi_n \equiv \frac{1}{\sqrt{n}} L'(X,\ \theta) = \frac{1}{\sqrt{n}} \sum l'(x_i,\ \theta)$$

es la suma de las variables independientes igualmente distribuidas, que tienen por media 0 y por varianza $I(\theta)$ (véase (2)), según el teorema central del límite $\xi \in \Phi_{0,I(\theta)}$. La representación (7) queda demostrada. Para demostrar (8) volvamos al lema (2). Este significa que existe un conjunto A, $\mathbf{P}_\theta(A) = 1$ tal que para $X_\infty \in A$, $n \to \infty$,

$$\sup_{|v|<\delta_n} \left| \frac{L'(X,\ \theta + v) - L'(X,\ \theta)}{nv} + I(\theta) \right| \to 0. \tag{13}$$

Además, en virtud del corolario 23.1 existen la sucesión $u_n \to \infty$, $u_n/\sqrt{n} \equiv \gamma_n \to 0$ (u_n debe satisfacer (23.9)) y el conjunto B, $\mathbf{P}_\theta(B) = 1$ tal que para $X_\infty \in B$, $n \to \infty$,

$$v^* = (\hat{\theta}^* - \theta) = o(\gamma_n). \tag{14}$$

Como la sucesión $\delta_n \to 0$ en (13) es arbitraria, para $X_\infty \in A \cap B$, $\mathbf{P}_\theta(A \cap B) = 1$, en virtud de (14) la relación (13) resultará justa en el punto $v = v^*$. Recordando que $L'(X,\ \theta + v^*) = L'(X,\ \hat{\theta}^*) = 0$, obtenemos para $X_\infty \in A \cap B$,

$$\left| I(\theta) - \frac{L'(X,\ \theta)}{n(\hat{\theta}^* - \theta)} \right| \to 0.$$

Esto significa que $\xi_n - I(\theta)u^* = u^*\varepsilon_n(X,\ \theta)$, y demuestra (8).

Haciendo uso de los mismos argumentos, se puede sustituir $u = u^* = v^*\sqrt{n} = (\hat{\theta}^* - \theta)\sqrt{n} = \dfrac{\xi_n}{I(\theta)}(1 + \varepsilon_n(X,\ \theta))$ en (12). Esto da

$$L(X,\ \hat{\theta}^*) - L(X,\ \theta) = \frac{\xi_n^2}{I(\theta)}(1 + \varepsilon_n(X,\ \theta))$$

y demuestra la primera parte de la relación (9). La convergencia de $\xi_n^2/I(\theta)$ hacia la distribución χ^2 con un grado de libertad se deduce de los teoremas de continuidad, ya que $\xi_n/\sqrt{I(\theta)} \in \Phi_{0,1}$.

La relación (10) se demuestra de un modo completamente análogo a (7) si se hace uso de la segunda afirmación del lema 2 y, basándose en ésta, se halla la representación para $L(X, \theta + u\sqrt{n}) - L(X, \hat{\theta}^*)$. ◁

Observación 2. En el lenguaje de las distribuciones, la primera afirmación del teorema 1 puede ser enunciada de la manera siguiente:

$$Y(u) \in \Phi_{-u^2 I(\theta)/2,\ u^2 I(\theta)}. \tag{15}$$

Anteriormente hemos señalado que la segunda condición (*RR*) (acerca de la existencia de $l''(x, \theta)$) no siempre es esencial para las afirmaciones que han de ser demostradas. El carácter no esencial de esta condición para la convergencia (15) se puede mostrar mediante los razonamientos siguientes. La magnitud

$$Y(u) = L\left(X, \theta + \frac{u}{\sqrt{n}}\right) - L(X, \theta) = \sum_{i=1}^{n}\left[l\left(x_i, \theta + \frac{u}{\sqrt{n}}\right) - l(x_i, \theta)\right]$$

es la suma de las magnitudes independientes igualmente distribuidas. Por eso, según el teorema central del límite para el esquema de series (los sumandos dependen de n y omitimos la verificación de las condiciones de Lindeberg)

$$Y(u) \in \Phi_{\alpha(u),\ \sigma^2(u)},$$

donde

$$\alpha(u) = \lim_{n\to\infty} n\mathbf{M}_\theta[l(x_1, \theta + u/\sqrt{n}) - l(x_1, \theta)] =$$

$$= \lim_{n\to\infty} n\mathbf{M}_\theta \ln \frac{f_{\theta+u/\sqrt{n}}(x_1)}{f_\theta(x_1)} = -u^2 \lim_{\Delta\to 0} \frac{Q_1(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta)}{\Delta^2} = -u^2 I(\theta)/2$$

(véase el teorema 21.2 y la observación 21.1). Luego

$$\sigma^2(u) = \lim_{n\to\infty} n\mathbf{M}_\theta[l(x_1, \theta + u/\sqrt{n}) - l(x_1, \theta)]^2 =$$

$$= u^2 \lim_{\Delta\to 0} \int \left[\frac{l(x, \theta + \Delta) - l(x, \theta)}{\Delta}\right]^2 f_\theta(x)\mu(dx) =$$

$$= u^2 \int (l'(x, \theta))^2 f_\theta(x)\mu(dx) = u^2 I(\theta).$$

Si al calcular $\alpha(u)$ y $\sigma^2(u)$ se utilizó el desarrollo $l(x, \theta + u/\sqrt{n})$ en serie con dos derivadas, obtendríamos el mismo resultado. Sin embargo, nos hemos cerciorado de que no es obligatorio hacer esto.

Concluyendo este párrafo, del teorema 1 obtendremos otro corolario útil que necesitaremos en adelante y que se refiere al comportamiento de las integrales de la relación de verosimilitud.

Teorema 2. *Supongamos que se cumplen las condiciones* (*RR*), *la función* $w(t)$ *satisface la condición*

$$|w(t)| \leqslant ce^{\alpha|t|^2}, \quad c < \infty, \quad \alpha = g/16$$

($g > 0$ está definido en el § 21) *y la función* $q(t)$ *es continua en el punto* $t = 0$ *y está limitada. Supongamos, además, que* Π *es cualquier medida en* $(R, \mathfrak{B})$, *tal que* $\int e^{-\alpha|u|^2/4}\Pi(du) < \infty$. *En este caso, si* θ *es un punto interior de* Θ y $X \in \mathbf{P}_\theta$,

$$J \equiv \int w(u^* - u)q(\theta + u/\sqrt{n})Z(u/\sqrt{n})\Pi(du) =$$

$$= e^{Y(u^*)}q(\theta)\left[\int w(u^* - u)e^{-\frac{1}{2}(u-u^*)^2 I(\theta)}\Pi(du) + \varepsilon_n(X, \theta)\right]. \qquad (16)$$

En particular, si Π *es la medida de Lebesgue,* $\Pi(du) = du$, *entonces*

$$J = \sqrt{\frac{2\pi}{I(\theta)}}\, e^{Y(u^*)}q(\theta)(\mathbf{M}w(\eta) + \varepsilon_n(X, \theta)),$$

$$\text{donde } \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0,\ \eta \in \Phi_{0,\, I^{-1}(\theta)}.$$

La afirmación (16) es muy natural, ya que el factor $q(\theta + u/\sqrt{n})$ es "casi constante" y la función $Z(u/\sqrt{n}) = e^{Y(u)}$ se aproxima, con una exactitud de hasta el factor constante, según el teorema 1, con una densidad de distribución normal.

Demostración. Para simplificar la notación nos limitaremos a examinar el caso cuando Π es la medida de Lebesgue. El paso al caso general no presenta ninguna dificultad.

Estimemos primeramente la parte de la integral (16) en la región $|u| > r$. Designémosla por $J(r)$. Como $f_\theta(X)/f_{\hat\theta^*}(X) \leqslant 1$, entonces, suponiendo, para abreviar, $Z = Z(u^*/\sqrt{n}) = e^{Y(u^*)}$, $t = \theta + u/\sqrt{n}$, obtenemos

$$Z^{-1}Z\left(\frac{u}{\sqrt{n}}\right) = \frac{f_t(X)}{f_{\hat\theta^*}(X)} \leqslant \left(\frac{f_t(X)}{f_{\hat\theta^*}(X)}\right)^{3/4} \leqslant Z^{3/4}\left(\frac{u}{\sqrt{n}}\right).$$

Por eso, en virtud de la desigualdad de Cauchy — Buniakovski, del teorema 23.1 y del corolario 23.1,

$$\mathbf{M}_\theta w(u^* - u)Z^{-1}Z(u/\sqrt{n}) \leqslant$$

$$\leqslant [\mathbf{M}_t w^2(\sqrt{n}(\hat\theta^* - t))\mathbf{M}_\theta Z^{1/2}(u/\sqrt{n})]^{1/2} \leqslant ce^{-u^2 g/4}.$$

Como máx $q(t) < \infty$, de aquí y del lema 23.1 hallamos

$$\mathbf{M}_\theta Z^{-1}J(r) \leqslant ce^{-gr^2/4}.$$

Haciendo uso de la desigualdad de Chébishev, obtenemos las estimaciones del mismo orden también para $\mathbf{P}_\theta(Z^{-1}J(r) > \delta)$. Por eso, si $r = r_n \to \infty$, de modo que

$$\sum e^{-r_n^2 g/4} < \infty, \qquad (17)$$

entonces, para $y \geqslant r_n$,

$$Z^{-1}J(y) \underset{\text{c.s.}}{\to} 0. \tag{18}$$

Elijamos $r_n = o(\sqrt{n})$ y examinemos la parte restante de la integral $V(y) = J - J(y)$ cuando $y = 2r_n$. Según el teorema 1,

$$Z^{-1}V(2r) = Z^{-1} \int\limits_{|u|<2r_n} q(\theta + u/\sqrt{n})w(u^* - u)Z(u/\sqrt{n})du =$$

$$= \int\limits_{|u|<2r_n} (q(\theta) + \varepsilon_n(u))w(u^* - u) \times$$

$$\times \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \varepsilon_n(X, \theta, u))\right\}du,$$

donde $|\varepsilon_n(u)| < \varepsilon_n \to 0$, $|\varepsilon_n(X, \theta, u)| \leqslant \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$ cuando $n \to \infty$. Por eso, en virtud de (18), es suficiente cerciorarse de la proximidad de las integrales

$$\int\limits_{|u|<2r_n} w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \varepsilon_n(X, \theta, u))\right\}du,$$

$$\sqrt{\frac{2\pi}{I(\theta)}}\ \mathbf{M}w(\eta) = \int w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\}du.$$

En virtud de (17) y del corolario 23.1 existe un conjunto A, $\mathbf{P}_\theta(A) = 1$ tal, que $|u^*| \leqslant r_n$ para $X_\infty \in A$ cuando todos $n = n(X_\infty)$ son bastante grandes. Como $I(\theta) \geqslant g$, $|u - u^*|^2 > u^2/2$ para $|u| > 2r_n$, $|u^*| < r_n$, entonces, en el conjunto A (véase el lema 23.1),

$$\int\limits_{|u|\geqslant 2r_n} w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\}du < ce^{-gr_n^2} \to 0.$$

Por eso nos queda estimar

$$\int\limits_{|u|<2r_n} w(u^* - u)\left|\exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \varepsilon_n(X, \theta, u))\right\} -\right.$$

$$\left. - \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\}\right|du \leqslant \int w(v)\left|\exp\left\{-\frac{1}{2}v^2 I(\theta) \times\right.\right.$$

$$\left.\left. - (1 + \varepsilon_n(X, \theta, v + u^*))\right\} - \exp\left\{-\frac{1}{2}v^2 I(\theta)\right\}\right|dv.$$

Pero esta integral converge en el conjunto AB hacia el cero, donde $B = \{X_\infty: \varepsilon_n(X, \theta) \to 0\}$, $\mathbf{P}_\theta(B) = 1$. Esto resulta de la convergencia a cero para cada v de la función subintegral y del hecho de que ésta es mayorada por la función sometida a integración. ◁

§ 25. Propiedades de las estimaciones de verosimilitud máxima. Normalidad asintótica. Optimación asintótica

Supongamos que $X \in \mathbf{P}_\theta$ y $\hat{\theta}^*$ es la e.v.m. Los resultados de los párrafos precedentes permiten describir por completo las propiedades asintóticas de θ^* cuando el volumen n de la muestra crece indefinidamente. Además, en este párrafo hemos establecido uno de los resultados centrales del capítulo presente, que consiste en que la e.v.m, al cumplirse las condiciones (RR), posee todas las propiedades posibles de optimación asintótica, que hemos examinado anteriormente, o sea, la estimación asintóticamente eficiente es, a la vez, asintóticamente bayesiana (para toda distribución a priori que tiene densidad) y asintóticamente minimax.

En este párrafo siempre supondremos, sin especificarlo complementariamente, que se cumplen las condiciones (RR).

1. Normalidad asintótica de la e.v.m.

Teorema 1. *La e.v.m. $\hat{\theta}^*$ es una estimación asintóticamente normal, con la particularidad de que la convergencia*

$$u^* = (\hat{\theta}^* - \theta)\sqrt{n} \in \Phi_{0,\, I^{-1}(\theta)} \tag{1}$$

tiene lugar junto con los momentos de cualquier orden, o sea, junto con (1), *para cualquier $k > 0$, se cumple*

$$\mathbf{M}_\theta (u^*)^k \to \mathbf{M}\eta^k, \quad \eta \in \Phi_{0,\, I^{-1}(\theta)}. \tag{2}$$

Además, para cualquier función continua $w(t)$ tal, que $|w(t)| < e^{gt^2/6}$ (véase (23.4)),

$$\mathbf{M}_\theta w(u^*) \to \mathbf{M}w(\eta), \quad \eta \in \Phi_{0,\, I^{-1}(\theta)}. \tag{3}$$

Demostración. En el teorema 24.1 hemos establecido que

$$u^* = (\hat{\theta}^* - \theta)\sqrt{n} = \frac{\xi_n}{I(\theta)}\,(1 + \varepsilon_n(X, \theta)), \tag{4}$$

donde $\varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$, $\xi_n = L'(X, \theta)/\sqrt{n} \in \Phi_{0, I(\theta)}$. Esto demuestra (1). Las relaciones (2) y (3) se obtienen de (1) y del teorema de continuidad para los momentos (véase el § 1.5), puesto que en virtud del corolario 23.2,

$$\mathbf{M}_\theta w^{6/5}(u^*) \leqslant \mathbf{M}_\theta \exp\left\{\frac{(u^*)^2 g}{5}\right\} < c < \infty. \quad \triangleleft$$

Observación 1. De (1) y (2) se deduce que $\hat{\theta}^*$ pertenece a la clase de estimaciones $\mathbf{K}_{\Phi,2}$, en la que la convergencia de $(\hat{\theta}^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2(\theta)}$ tiene lugar junto con la convergencia de $\mathbf{M}_\theta(\theta^* - \theta)^2 \to \sigma^2(\theta)$ de los primeros momentos. Como ya hemos señalado en el § 8, en esta clase, el enfoque asintótico de la comparación de las estimaciones coincide, de hecho, con el enfoque estándar.

Observación 2. La relación (4) también permite describir exactamente las "desviaciones máximas" de $(\hat{\theta}^* - \theta)\sqrt{n}$ cuando $n \to \infty$. Pues, se sabe (véanse [61] y [84]) que las sumas normalizadas ξ_n de las magnitudes independientes igualmente distribuidas, que tienen por media el cero y por varianza $I(\theta)$, satisfacen la ley de logaritmo reiterado, en virtud de la cual

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{|\xi_n|}{\sqrt{2I(\theta)\ln\ln n}} = 1\right) = 1.$$

En vista de que en (4) $\limsup_{n\to\infty} \varepsilon_n(X, \theta) = 0$ c.s., obtenemos que

$$\mathbf{P}_\theta\left(\limsup_{n\to\infty} \frac{|\hat{\theta}^* - \theta|\sqrt{nI(\theta)}}{\sqrt{2\ln\ln N}} = 1\right) = 1.$$

Determinemos ahora, en calidad de corolarios del teorema 2, algunas propiedades de la e.v.m. relacionadas con la optimación asintótica.

2. Eficacia asintótica. En el § 16 hemos introducido el estudio de la clase $\tilde{K}_0$ de estimaciones asintóticamente no desplazadas, o sea, de estimaciones θ^* cuyo desplazamiento $b(\theta) = \mathbf{M}_\theta\theta^* - \theta$ posee las propiedades

$$b(\theta) = o(1/\sqrt{n},\ b'(\theta) = o(1). \tag{5}$$

En el § 20 hemos expuesto las ideas según las cuales, en búsqueda de las estimaciones asintóticamente eficientes "en total", es posible limitarse a la clase $\tilde{K}_0$.

Establezcamos ahora el hecho siguiente.

Corolario 1. $\theta^* \in \tilde{K}_0$.

Demostración. La primera de las relaciones (5) resulta de (2) cuando $k = 1$. Para demostrar la segunda señalemos que (véase el § 16)

$$1 + b'(\theta) = \mathbf{M}_\theta\hat{\theta}^*L'(X, \theta) = \mathbf{M}_\theta(\hat{\theta}^* - \theta)L'(X, \theta) =$$

$$= \mathbf{M}_\theta((\hat{\theta}^* - \theta)\sqrt{n}\,\xi_n) = \mathbf{M}_\theta\frac{\xi_n^2}{I(\theta)}(1 + \varepsilon_n(X, \theta)),$$

$$\varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0.$$

Si aquí es cierto el teorema de continuidad para los momentos, entonces obtenemos la relación requerida $1 + b'(\theta) \to 1$ o, que es lo mismo, $b'(\theta) \to 0$. Para establecer la validez de este teorema en nuestro caso, es

suficiente cerciorarse (véase el § 1.5) de que

$$\mathbf{M}_\theta|(\hat{\theta}^* - \theta)\sqrt{n}\,\xi_n|^{3/2} < c < \infty, \tag{6}$$

donde c no depende de n. Hagamos uso de la desigualdad de Hölder

$$\mathbf{M}|\xi\eta|^r \leqslant (\mathbf{M}|\xi|^{pr})^{1/p}(\mathbf{M}|\eta|^{qr})^{1/q}, \quad p > 0, \quad q > 0, \quad \frac{1}{p} + \frac{1}{q} = 1$$

para $r = 3/2$, $p = 4$, $q = 4/3$. Entonces obtenemos, para el primer miembro de (6), la estimación $(\mathbf{M}_\theta[(\hat{\theta}^* - \theta)\sqrt{n}]^6)^{1/4}(\mathbf{M}\xi_n^2)^{3/4}$, que, en virtud de (2), nos da la desigualdad deseada. ◁

El corolario siguiente, debido a su importancia, lo enunciaremos en forma de teorema.

Teorema 2. *La e.v.m. $\hat{\theta}^*$ es una estimación asintóticamente R-eficiente. Además, $\hat{\theta}^*$ es asintóticamente eficiente en $\tilde{K}_0$.*

Demostración. El hecho de que $\hat{\theta}^*$ es una estimación asintóticamente R-eficiente se desprende directamente de la definición 16.1 y del hecho de que

$$\mathbf{M}_\theta(\hat{\theta}^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

La eficacia asintótica en $\tilde{K}_0$ se deduce del teorema 16.3. ◁

El teorema 2, junto con las observaciones referentes al teorema 16.3, significa que, al cumplirse las condiciones (RR), cualquier estimación asintóticamente eficiente en $\tilde{K}_0$ será una estimación asintóticamente R-eficiente.

Anotemos que la contracción del conjunto de las estimaciones examinadas, hasta $\tilde{K}_0$, no es la única contracción, ni mucho menos, con la que $\hat{\theta}^*$ se vuelve asintóticamente eficiente.

Indiquemos otra contracción relacionada en este caso con la propiedad de θ de ser mediana asintótica de la distribución de las estimaciones asintóticamente normales, o sea, con la propiedad

$$\mathbf{P}_\theta(\hat{\theta}^* > \theta) \to 1/2 \tag{7}$$

cuando $n \to \infty$.

Designemos por $\tilde{K}^\circ$ la clase de estimaciones θ^* para las cuales (7) se cumple uniformemente respecto a θ. La clase $\tilde{K}^\circ$ podría llamarse clase de *estimaciones asintóticamente centrales.*

Teorema 3. *La e.v.m. $\hat{\theta}^* \in \tilde{K}^\circ$ es precisamente una estimación asintóticamente eficiente en la clase $\tilde{K}^\circ$.*

Aplazaremos la demostración de este teorema hasta el § 3.3.

3. Carácter asintóticamente bayesiano de la e.v.m. En este apartado, por doquier se suponga la existencia de la densidad $q(t)$ de la distribución a

priori $\mathbf{Q}$ respecto a la medida de Lebesgue en Θ, supondremos también, sin especificarlo complementariamente, que la densidad es integrable según Riemann, así que se satisfarán las condiciones del teorema 20.5.

Teorema 4. *La e.v.m. $\hat{\theta}^*$ es una estimación asintóticamente R-bayesiana. Si $\mathbf{Q}$ es una distribución arbitraria a priori que tiene una densidad $q(t)$ respecto a la medida de Lebesgue, entonces $\hat{\theta}^*$ también es una estimación asintóticamente bayesiana que* corresponde a la distribución $\mathbf{Q}$.

Demostración. El carácter asintóticamente R-bayesiano de la e.v.m. se deduce de las relaciones

$$\lim_{n\to\infty} \mathbf{M}[\sqrt{n}(\hat{\theta}^* - \theta)]^2 = \lim_{n\to\infty} \mathbf{M}\mathbf{M}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2 =$$

$$= \mathbf{M} \lim_{n\to\infty} \mathbf{M}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2 = \mathbf{M}I^{-1}(\theta) = J.$$

Aquí el paso límite bajo el signo de la esperanza matemática es legítimo según el teorema de la convergencia mayorada, ya que, en virtud de 23.2, el valor de $\mathbf{M}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2$ está uniformemente limitado por la constante que no depende de n ni de θ.

El carácter asintóticamente bayesiano se deduce del corolario 20.1. ◁

De las observaciones referentes al corolario 20.1 y del teorema 4 resulta que cualquier estimación asintóticamente bayesiana es asintóticamente R-bayesiana.

La afirmación del teorema 4 puede ser amplificada. Resulta que la e.v.m. y la estimación bayesiana "casi" coinciden para cualquier densidad a priori q.

Teorema 5.

$$\mathbf{M}n(\hat{\theta}^* - \theta_{\mathbf{Q}}^*)^2 \to 0, \quad (\theta_{\mathbf{Q}}^* - \hat{\theta}^*)\sqrt{n} \underset{p}{\to} 0,$$

donde $\theta_{\mathbf{Q}}^*$ es la estimación bayesiana que corresponde a la distribución $\mathbf{Q}$, y la convergencia en probabilidad se entiende respecto a la distribución compatible de X y θ en $\mathscr{X}^n \times \Theta$.

El teorema 5 se desprende directamente del corolario 20.2. Su afirmación es equivalente a que para casi todos t

$$\mathbf{M}_t n(\hat{\theta}^* - \theta_{\mathbf{Q}}^*)^2 \to 0.$$

Es posible la amplificación ulterior de la afirmación enunciada.

Teorema 6. *Sea θ un punto interior arbitrario Θ, $X \in \mathbf{P}_\theta$. Sea, luego, $q(t)$ una densidad arbitraria, continua y positiva dentro de Θ, de la distribución a priori. Entonces $\sqrt{n}(\hat{\theta}^* - \theta_{\mathbf{Q}}^*) \underset{\text{c.s.}}{\to} 0$.*

La demostración de deduce del teorema 2 del párrafo precendente. En efecto, $\theta_Q^* - \hat{\theta}^* = \dfrac{\int (t - \hat{\theta}^*) q(t) f_t(X) dt}{\int q(t) f_t(X) dt}$. Sustituyendo las variables $t = \theta + u/\sqrt{n}$ y dividiendo por $f_\theta(X)$ el numerador y denominador en esta expresión, obtenemos

$$\theta_Q^* - \hat{\theta}^* = \frac{\int (u - u^*) q(\theta + u/\sqrt{n}) Z(u/\sqrt{n}) du}{\sqrt{n} \int q(\theta + u/\sqrt{n}) Z(u/\sqrt{n}) du}.$$

Ahora es necesario hacer uso del teorema 24.2 para $w(t) = t$ y $w(t) = 1$. Como en el primer caso $\mathbf{M} w(\eta) = \mathbf{M}\eta = 0$, entonces obtenemos

$$\theta_Q^* - \hat{\theta}^* = \varepsilon_n(X, \theta)/\sqrt{n}, \quad \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0. \ \triangleleft$$

4. Carácter asintóticamente minimax de la e.v.m.

Teorema 7. *La e.v.m. es una estimación asintóticamente minimax.*

Este teorema se deduce directamente del corolario 20.3 y de la afirmación siguiente.

Lema 1.

$$\lim_{n \to \infty} \sup_{\theta \in \Gamma} \mathbf{M}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta \in \Gamma} I^{-1}(\theta),$$

donde Γ *es cualquier trazado dentro de* Θ.

El lema 1 se desprende de la convergencia (2) uniforme en θ. La uniformidad será demostrada en el § 29 (véase el apartado 29.3).

§ 26*. Cálculo aproximado de las estimaciones de verosimilitud máxima

Hemos visto que en los problemas de estimación de los parámetros revisten el máximo interés las estimaciones eficientes y asintóticamente eficientes y, en particular, las e.v.m. Surge la cuestión acerca de la determinación práctica de tales estimaciones. En los problemas reales, la búsqueda del valor exacto de la e.v.m. $\hat{\theta}^*$ puede presentar grandes dificultades. Esto se refiere, sobre todo, a las distribuciones que no tienen estadísticas suficientes relativamente sencillas.

Por otro lado, la determinación de cualquier estimación asintóticamente normal θ^* no provoca, por regla general, dificultades.

Aquí mostraremos un método de construcción de la estimación θ_1^*, asintóticamente equivalente a la e.v.m. $\hat{\theta}^*$ (y, por consiguiente, a la asintóticamente eficiente), el cual se basa en el método de Newton para cálculos

aproximados y en la utilización de la estimación asintóticamente normal θ^*. Pongamos

$$U(t) = t - L'(X, t)\cdot(L''(X, t))^{-1}, \quad t \in \Theta,$$
$$U_1(t) = t + L'(X, t)\cdot(nI(t))^{-1}, \quad t \in \Theta.$$

Teorema 1. *Supongamos que se cumplen las condiciones (RR), $X \in \mathbf{P}_\theta$ y que θ^* es cualquier estimación asintóticamente normal*

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\ \sigma^2(\theta)}.$$

En este caso la estimación $\theta_1^* = U(\theta^*)$ (o bien $\theta_1^* = U_1(\theta^*)$) será asintóticamente equivalente a $\hat\theta^*$, o sea,

$$(\theta_1^* - \hat\theta^*)\sqrt{n} \underset{P_\theta}{\to} 0.$$

La demostración del teorema se apoyará en el lema siguiente.

Lema 1. *Supongamos que se cumplen las condiciones (RR), $X \in \mathbf{P}_\theta$, y que $\delta_n > 0$ es una sucesión arbitraria convergente a cero. En este caso, si θ_n es tal que $|\theta_n - \theta| \leqslant \delta_n$,*

$$U(\theta_n) - \hat\theta^* = (\theta_n - \hat\theta^*)\varepsilon_n(\theta_n, \theta, X),$$

donde $\bar\varepsilon_n \equiv \max\limits_{\theta_n:|\theta_n-\theta|\leqslant\delta_n} |\varepsilon_n(\theta_n, \theta, X)| \underset{P_\theta}{\to} 0.$

Esa misma afirmación será válida si en vez de U utilizamos la función U_1.

Con otras palabras, si se hace uso del método de aproximaciones sucesivas hacia $\hat\theta^*$ y se pone $\theta_0^* = \theta_n$, $\theta_1^* = U(\theta_0^*)$ (o bien $\theta_1^* = U_1(\theta_0^*)$), entonces $\theta_1^* - \hat\theta^* = o(\theta_0^* - \hat\theta^*)$, así que la aproximación θ_1^* es mucho mejor que θ_0^*.

Demostración. De las investigaciones de § 24 y de la continuidad de L'' se deduce (véase, por ejemplo, el lema 24.1) que

$$L'(X, \theta_n) = (\theta_n - \hat\theta^*)L''(X, \tilde\theta), \quad L''(X, \tilde\theta) = n(I(\theta) + \varepsilon_n'(\theta_n, \theta, X)),$$

donde $\tilde\theta \in [\theta_n, \hat\theta^*]$, $\max\limits_{\theta_n:|\theta_n-\theta|\leqslant\delta_n} \varepsilon_n'(\theta_n, \theta, X) \underset{P_\theta}{\to} 0$ para cualquier sucesión $\delta_n \to$ $\to 0$. Luego,

$$L''(X, \theta_n) = n(I(\theta) + \varepsilon_n''),$$
$$(I(\theta) + \varepsilon_n')(I(\theta) + \varepsilon_n'')^{-1} = 1 + \varepsilon_n,$$

donde ε_n'', ε_n poseen la misma propiedad que ε_n'. Por consiguiente,

$$U(\theta_n) - \hat\theta^* = \theta_n - \theta^* - L'(X, \theta_n)(L''(X, \theta_n))^{-1} =$$
$$= \theta_n - \hat\theta^* - (\theta_n - \theta^*)(1 + \varepsilon_n) = (\theta_n - \hat\theta^*)\varepsilon_n.$$

La demostración para la función U_1 se realiza exactamente igual. ◁

Demostración del teorema 1. Elijamos cualquier $\delta_n \to 0$ tal, que $\delta_n\sqrt{n} \to \infty$, y representemos $(\theta_1^* - \hat{\theta}^*)\sqrt{n}$ en la forma

$$(U(\theta^*) - \hat{\theta}^*)\sqrt{n} = \sqrt{n}(\theta^* - \hat{\theta}^*)\varepsilon_n(\theta^*, \theta, X)I_{(|\theta^* - \theta| \leqslant \delta_n)} + r_n,$$

donde $r_n \neq 0$ únicamente en el conjunto $B_n = \{X: |\theta^* - \theta| > \delta_n\}$ y, en virtud del lema 1,

$$\bar{\varepsilon}_n = \max_{|t-\theta| \leqslant \delta_n} \varepsilon_n(t, \theta, X) \underset{P_\theta}{\to} 0.$$

Como, además, $\mathbf{P}_\theta(B_n) \to 0$, de aquí se deduce que

$$|\theta_1^* - \hat{\theta}^*|\sqrt{n} \leqslant \sqrt{n}|\theta^* - \theta|\bar{\varepsilon}_n + \sqrt{n}|\hat{\theta}^* - \theta|\bar{\varepsilon}_n + r_n \underset{P_\theta}{\to} 0. \quad \triangleleft$$

El teorema 1 muestra que el método de aproximaciones sucesivas, partiendo de cualquier estimación asintóticamente normal, nos lleva en 1 paso al punto θ^*, con una exactitud de hasta los valores de $o(1/\sqrt{n})$.

Si se exige la existencia de las terceras derivadas continuas $l'''(x, \theta)$, entonces también se puede comenzar de puntos más lejos, que distan de θ, digamos, a la magnitud de $o(n^{-1/4})$. En este caso, al igual que en las condiciones del teorema 1, en 1 paso resultaremos en el $o(1/\sqrt{n})$-entorno del punto $\hat{\theta}^*$. En efecto,

$$L'(X, t) = (t - \hat{\theta}^*)L''(X, \hat{\theta}^*) + \frac{(t - \hat{\theta}^*)^2}{2} L'''(X, \theta') =$$

$$= (t - \hat{\theta}^*)L''(X, t) + \frac{3}{2}(t - \hat{\theta}^*)^2 L'''(X, \theta''),$$

donde θ' y θ'' están comprendidos entre t y $\hat{\theta}^*$. Por eso

$$U(\theta_n) - \hat{\theta}^* = \theta_n - \hat{\theta}^* - L'(X, \theta_n)(L''(X, \theta_n))^{-1} =$$

$$= \frac{3}{2}(\theta_n - \hat{\theta}^*)^2(I(\theta) + \varepsilon_n), \quad \sqrt{n}(U(\theta_n) - \hat{\theta}^*) \underset{P_\theta}{\to} 0$$

si $|\theta_n - \theta| = o(n^{-1/4})$. $\triangleleft$

Ejemplo 1. *Clasificación de las partículas.* Examinemos una fuente que emite partículas de dos tipos: con probabilidad p, partículas del tipo A; y con probabilidad 1, p partículas del tipo B. La energía de las partículas es aleatoria y tiene una densidad de $f_1(x)$ para las partículas del tipo A, y de $f_2(x)$ para las del tipo B. Las funciones $f_i(x)$ son conocidas. Han sido registradas n partículas con energías $x_1, \ldots, x_n$. ¿A qué es igual la probabilidad p? Aquí la función de verosimilitud es igual a

$$f_p(X) = \prod_{i=1}^{n} (pf_1(x_i) + (1 - p)f_2(x_i)),$$

así que

$$L'(X, p) = \sum_{i=1}^{n} \frac{f_1(x_i) - f_2(x_i)}{pf_1(x_i) + (1 - p)f_2(x_i)}. \tag{1}$$

Vemos que la búsqueda de la e.v.m. $\hat{p}^*$ conduce a la ecuación $L' = 0$ de grado $n - 1$ respecto a p, la cual se resuelve, para grandes n, con mucha dificultad. Hagamos uso del teorema 1. Para eso necesitamos cualquier estimación asintóticamente normal p^*. Supongamos que $\int (F_1 - F_2)^2 dx < \infty$, donde $F_i(x) = \int_{-\infty}^{x} f_i(t)dt$, y examinemos el enfoque natural siguiente. Definamos p^* como valor que minimiza

$$\int (F_n^*(x) - F(x))^2 dx, \quad F(x) = pF_1(x) + (1 - p)F_2(x). \tag{2}$$

Igualando a cero la derivada de (2), obtenemos $\int (F_n^* - F)(F_1 - F_2)dx = 0$,

$$p^* = \frac{\int (F_n^* - F_2)(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}.$$

Es fácil notar que $\mathbf{M}p^* = p$ y que

$$(p^* - p)\sqrt{n} = \frac{\int (F_n^* - F)\sqrt{n}(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}. \tag{3}$$

De los resultados de los §§ 1.6—1.8 se deduce que p^* es una estimación asintóticamente normal y que la distribución límite (3) coincide con la distribución

$$\frac{\int w\,°(F(x))(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}.$$

Por lo tanto, en virtud del teorema 1 la estimación

$$p_1^* = p^* - L'(X, p^*)(L''(X, p^*))^{-1},$$

donde L' está definida en (1),

$$L'' = -\sum \frac{(f_1(x_i) - f_2(x_i))^2}{(pf_1(x_i) + (1 - p)f_2(x_i))^2},$$

será asintóticamente equivalente a la e.v.m. $\hat{p}^*$. El coeficiente de dispersión p_1^* será determinado por la información

$$I(p) = \int \frac{(f_1(x) - f_2(x))^2}{pf_1(x) + (1 - p)f_2(x)} dx$$

y será menor que el coeficiente de dispersión p^*.

Ejemplo 2. Le proponemos al lector que halle, de ese mismo modo, la aproximación para la e.v.m. del parámetro α de la distribución de Cauchy

$\mathbf{K}_{\alpha,1}$ que tiene una densidad de

$$k_{\alpha,1}(x) = \frac{1}{\pi(1 + (x - \alpha)^2)}.$$

En calidad de estimación asintóticamente normal "previa" se puede tomar la mediana muestral ζ^* (véase el § 2 ó los §§ 1.3 y 1.8 Aquí no se puede tomar la estimación $\alpha^* = \bar{x}$, ya que $\mathbf{M}_\alpha \alpha^*$ no existe). La estimación

$$\alpha_1^* = \zeta^* - L'(X, \zeta^*)(L''(X, \zeta^*))^{-1},$$

donde

$$L'(X, \alpha) = -2 \sum \frac{x_i - \alpha}{1 + (x_i - \alpha)^2},$$

$$L''(X, \alpha) = 2 \sum \frac{1 - (x_i - \alpha)^2}{(1 + (x_i - \alpha)^2)^2},$$

será asintóticamente equivalente a la e.v.m. $\hat{\alpha}^*$. Como

$$I(\alpha) = \int \frac{(k'_{\alpha,1}(x))^2}{k_{\alpha,1}(x)}\, dx = \frac{4}{\pi} \int \frac{x^2}{(1 + x^2)^3}\, dx = \frac{1}{2},$$

los coeficientes de dispersión ζ^* y α_1^* serán iguales respectivamente (véase el § 2) a

$$\frac{1}{2k_{\alpha,1}(\alpha)} = \frac{\pi}{2}, \quad I^{-1/2}(\alpha) = \sqrt{2}, \quad \frac{\pi}{2} > \sqrt{2}.$$

Ejemplo 3. La sangre de cada persona pertenece a uno de los cuatro grupos que designamos por 0 (cero), A, B y AB. El heredamiento de los grupos de sangre es controlado por tres genes: A, B y 0, además, el gene 0 es "deprimido" por los genes A y B. Por eso, si p, q y $r = 1 - p - q$ designan las probabilidades de que aparezcan los genes A, B y 0, las probabilidades de aparición de los grupos de sangre corresponderán a las siguientes magnitudes:

Tabla 1

i (número de grupo)	Grupo	Combinaciones que dan este grupo	Probabilidades
1	0	00	r^2
2	A	AA, A0	$p^2 + 2pr$
3	B	BB, B0	$q^2 + 2qr$
4	AB	AB	$2pq$

Tabla 2

	i			
	1	2	3	4
$p_i(\theta)$	r^2	$p(p + 2r)$	$q(q + 2r)$	$2pq$
$\frac{\partial p_i(\theta)}{\partial p}$	$-2r$	$2r$	$-2q$	$2q$
$\frac{\partial p_i(\theta)}{\partial q}$	$-2r$	$-2p$	$2r$	$2p$

Sean ν_1, ν_2, ν_3, ν_4 las frecuencias de aparición de los grupos de sangre respectivos en la población sujeta a investigación, con un total de n personas. ¿Cómo hallar la e.v.m. par p y q? En nuestro caso las probabilidades $p_i(\theta)$, $\theta = (p, q)$ de aparición del i-ésimo grupo de sangre y sus derivadas parciales respecto a p y q se muestran en la tabla 2.

Por eso para la función logarítmica de verosimilitud $L(X, \theta) = \sum_{i=1}^{4} \nu_i \ln p_i(\theta)$ obtenemos

$$\frac{\partial L}{\partial p} = \sum \frac{\nu_i}{p_i} \frac{\partial p_i}{\partial p} = -\frac{2\nu_1}{r} + \frac{2r\nu_2}{p(p+2r)} - \frac{2\nu_3}{q+2r} + \frac{\nu_4}{p},$$
$$\frac{\partial L}{\partial q} = \sum \frac{\nu_i}{p_i} \frac{\partial p_i}{\partial q} = -\frac{2\nu_1}{r} - \frac{2\nu_2}{p+2r} + \frac{2r\nu_3}{q(q+2r)} + \frac{\nu_4}{q}. \quad (4)$$

Igualando a cero estas derivadas, llegaremos al sistema de dos ecuaciones para θ^* de cuarto orden. La resolución de tal sistema presenta dificultades técnicas. Por eso es más simple hacer uso del teorema 1. Para esto notemos que son válidas las igualdades

$$p_1 = r^2, \quad p_1 + p_2 = (p + r)^2, \quad p_1 + p_3 = (q + r)^2. \quad (5)$$

Las estimaciones eficientes para p_i son iguales a $p_i^* = \nu_i/n$. Sustituyendo en (5) estas estimaciones y resolviendo las ecuaciones obtenidas, tenemos

$$p^* = \sqrt{p_1^* + p_2^*} - \sqrt{p_1^*}, \quad q^* = \sqrt{p_1^* + p_3^*} - \sqrt{p_1^*}.$$

Como p_i^* es la estimación asintóticamente normal de p_i (o sea, $(p_i^* - p_i)\sqrt{n} \in \Phi_{0,\, p_i(1-p_i)}$), en virtud de los teoremas del § 1.5, p^* y q^* también serán las estimaciones asintóticamente normales para p y q.

Para valerse del teorema 1 sólo queda calcular la matriz $(L''(X, \theta^*))^{-1}$ o matriz $(nI(\theta^*))^{-1}$, $\theta^* = (p^*, q^*)$.

Citemos el ejemplo de una muestra real X obtenida como resultado del examen de $n = 353$ personas.

La distribución de la gente por grupos de sangre se da en la tabla 3.

Tabla 3

	0	*A*	*B*	*AB*	*Total*
ν_i	121	120	79	33	353
p_i^*	0,343	0,340	0,224	0,093	1

Tabla 3A

	0	*A*	*B*	*AB*
$p_i(\theta^*)$	0,351	0,343	0,226	0,080
$p_i(\theta_1^*)$	0,337	0,347	0,231	0,085

De esta tabla se deduce $p^* = 0{,}241$, $q^* = 0{,}167$, $r^* = 1 - p^* - q^* = 0{,}592$. Con ayuda de la tabla 2, para los elementos de la matriz $I(\theta)$, cuando

$\theta = \theta^*$; obtenemos

$$\sum \left(\frac{\partial p_i(\theta)}{\partial p}\right)^2 \frac{1}{p_i(\theta)} = 4 + \frac{4r^2}{p(p+2r)} + \frac{4q}{q+2r} + \frac{2q}{p} = 9{,}970,$$

$$\sum \left(\frac{\partial p_i(\theta)}{\partial p}\right)^2 \frac{1}{p_i(\theta)} = 4 + \frac{4p}{p+2r} + \frac{4r^2}{q(q+2r)} + \frac{2p}{q} = 13{,}761,$$

$$\sum \frac{\partial p_i(\theta)}{\partial p} \cdot \frac{\partial p_i(\theta)}{\partial q} \cdot \frac{1}{p_i(\theta)} = 4 - \frac{4r}{p+2r} - \frac{4r}{q+2r} + 2 = 2{,}585.$$

De aquí hallamos $|I(\theta^*)| = 130{,}512$.

$$I^{-1}(\theta^*) = \begin{vmatrix} 0{,}105 & -0{,}020 \\ -0{,}020 & 0{,}076 \end{vmatrix}.$$

De las fórmulas para $\frac{\partial L}{\partial p}$ y $\frac{\partial L}{\partial q}$ (véase (4)) obtenemos

$$L'(\theta^*, X) = (25{,}443,\ 34{,}161), \tag{6}$$

así que para la segunda aproximación de θ_1^* tenemos

$$\theta_1^* = \theta^* + \frac{1}{n} L'(\theta^*, X) I^{-1}(\theta^*) = (0{,}246,\ 0{,}173). \tag{7}$$

Esto nos da, para completar la tabla 3, las estimaciones expuestas en la tabla 3A.

La aplicación de una iteración más, en forma de (7), ya no modifica la estimación θ_1^* (dentro de los límites de la exactitud que utilizamos), ya que

$$L'(\theta_1^*, X) = (-0{,}076,\ -0{,}167)$$

(compárese con (6)), así que la tercera aproximación para $\hat{\theta}^*$ y todas las aproximaciones siguientes coincidirán con θ_1^*.

§ 27*. Propiedades de las estimaciones de verosimilitud máxima al faltar las condiciones de regularidad. Conciliabilidad

Este párrafo, al igual que el § 22, no entra en el curso principal de exposición y está dedicado al estudio de un caso irregular. Aquí nos limitaremos a demostrar la conciliabilidad fuerte de la e.v.m. en condiciones muy débiles respecto a $f_t(x)$, las cuales no suponen el cumplimiento de las condiciones (RR) o (R). Un estudio más detallado de las propiedades de la e.v.m. y de la relación de verosimilitud en el caso irregular véase en [48].

En todo el párrafo supondremos que se cumplen las condiciones (A_p),

(A_c) y (A_0) y designaremos la distancia de Kullback-Leibler $\varrho_1(\mathbf{P}_\theta, \mathbf{P}_t)$ por

$$\varrho(\theta, t) = \int \ln \frac{f_\theta(x)}{f_t(x)} \cdot f_\theta(x)\mu(dx).$$

Sabemos que $\varrho(\theta, t) > 0$ para $t \neq \theta$ si se cumple la condición (A_0).

Evidentemente, la condición (A_0) es necesaria para la conciliabilidad de la e.v.m., o sea, para la convergencia de $\hat{\theta}^* \underset{P_\theta}{\to} \theta$. Si, por ejemplo, $\varrho(\theta, t_0) = 0$ cuando $t_0 \neq \theta$, entonces los puntos θ y t_0 serán simplemente indistinguibles, las distribuciones $\mathbf{P}_\theta$ y $\mathbf{P}_{t_0}$ coincidirán y cualquiera que sea el lugar de convergencia de la e.v.m. $\hat{\theta}^*$, ésta no podrá ser conciliable si $X \in \mathbf{P}_\theta$ o si $X \in \mathbf{P}_{t_0}$.

La siguiente variante de la condición (A_0) se puede llamar uniforme (θ ha sido registrado):

$(\overline{A_0})$ *Para cualquier* $\delta = \varepsilon(\delta) > 0$

$$\inf_{t:\ |t-\theta| \geqslant \delta} \varrho(\theta, t) > \varepsilon$$

con cierto $\varepsilon > 0$.

Es evidente que $(\overline{A_0})$ será el corolario de (A_0), (A_c) y de la continuidad de $\varrho(\theta, t)$. Por consiguiente, en estas condiciones, la condición $(\overline{A_0})$ también será necesaria.

Examinemos ahora la siguiente amplificación de la condición $(\overline{A_0})$. Designemos

$$f_t^\Delta(x) = \sup_{|u| \leqslant \Delta} f_{t+u}(x).$$

(A_0^Δ). *Para cualquier* $\delta > 0$ *existe* $\Delta = \Delta(\delta) > 0$ *tal, que para todos* t, $|t - \theta| > \delta$,

$$\int \ln \frac{f_t^\Delta(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < -\varepsilon \tag{1}$$

con cierto $\varepsilon > 0$.

Esta condición resulta suficiente para la conciliabilidad fuerte de la e.v.m. La misma es parecida a la condición $(\overline{A_0})$ y en este sentido se asemeja a la condición necesaria. Una sola condición (A_0) no es suficiente para la conciliabilidad de la e.v.m. (véase la observación 1).

Teorema 1. *Si se cumple la condición* (A_0^Δ), *entonces la e.v.m.* $\hat{\theta}^*$ *es fuertemente conciliable.*

Demostración. La e.v.m. $\hat{\theta}^*$ es el punto t en el que se alcanza el máximo de la función $\psi(t, \theta, \mathbf{P}_n^*)$, donde

$$\psi(\theta, t, \mathbf{P}) = \int \ln \frac{f_t(x)}{f_\theta(x)} \mathbf{P}(dx).$$

Como $\psi(\theta, \hat{\theta}^*, \mathbf{P}_n^*) \geqslant \psi(\theta, \theta, \mathbf{P}_n^*) = 0$, para demostrar el teorema es suficiente convencerse de que con $\mathbf{P}_\theta$-probabilidad igual a 1,

$$\lim_{n\to\infty} \sup_{|t-\theta|\geqslant\delta} \psi(\theta, t, \mathbf{P}_n^*) < -\varepsilon$$

con cierto $\varepsilon > 0$. (Esto precisamente significará que para c.t. $X_\infty \in \mathbf{P}_\theta$, a partir de cierto $n = n(X_\infty) < \infty$, se cumple $|\hat{\theta}^* - \theta| < \delta$). Supongamos que se ha registrado δ y que Δ satisface la condición (1). Recubramos el conjunto $\Theta \setminus [\theta - \delta, \quad \theta + \delta]$ con segmentos $\Delta_k = \{t: \quad |t - t_k| \leqslant \Delta\}, \quad k = 1, \ldots, N < \infty$, donde $t_k \in \Theta$, $t_k \notin [\theta - \delta, \theta + \delta]$. En este caso, según la ley fuerte de los grandes números,

$$\sup_{|t-\theta|\geqslant\delta} \psi(\theta, t, \mathbf{P}_n^*) \leqslant \max_k \sup_{t\in\Delta_k} \psi(\theta, t, \mathbf{P}_n^*) \leqslant$$

$$\leqslant \max_k \frac{1}{n} \sum_{i=1}^{n} \sup_{t\in\Delta_k} \ln \frac{f_t(x_i)}{f_\theta(x_i)} \underset{\text{c.s.}}{\to} \max_k \mathbf{M}_\theta \ln \frac{f_{t_k}^{\Delta}(x_1)}{f_\theta(x_1)} < -\varepsilon. \quad \triangleleft$$

Observación 1. Como ya hemos señalado, una sola condición $(\overline{A_0})$ no es suficiente para la conciliabilidad de $\hat{\theta}^*$. Para convencerse de esto examinemos el ejemplo siguiente. Sea $\Theta = [0, 1]$, $\mathbf{P}_\theta = \mathrm{U}_{\theta, 1+\theta}$ cuando $0 \leqslant \theta \leqslant 1/2$ y cuando $\theta = 1$. Cuando $1 > \theta > 1/2$, la distribución $\mathbf{P}_\theta$ tiene una densidad de $f_\theta(x) = 1/\theta$ cuando $1 - \theta < x < 1$. Supongamos ahora que $X \in \mathbf{P}_0 = \mathrm{U}_{0,1}$. En este caso la condición $(\overline{A_0})$ se cumple, ya que $\varrho(0, t) = -\infty$ cuando $t \neq 0$. Al mismo tiempo es fácil ver que $f_t(X) > 1$ cuando $t \in (1 - x_{(1)}, 1)$ y que $\hat{\theta}^* = 1 - x_{(1)} \underset{\text{c.s.}}{\to} 1$.

Las condiciones (A_0^{Δ}) pueden ser representadas de manera equivalente en una forma algo distinta. Designemos $f_t^\circ(x) = \lim\limits_{u\to t} \sup f_u(x)$.

Teorema 2. *La condición (A_0^{Δ}) es equivalente al cumplimiento simultáneo de las dos condiciones siguientes*

$(A\delta)$. Para todos $t \neq \theta$

$$\int \ln \frac{f_t^\circ(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < 0.$$

(J). Para todos t y cierto $\Delta > 0$

$$\int \ln \frac{f_t^{\Delta}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < \infty.$$

La condición (J), al igual que $(A\delta)$, (A_0^{Δ}), significa la integrabilidad de las partes positivas de las funciones subintegrales. Tales funciones es natural llamarlas *integrables superiormente.*

En virtud de (A_c), la condición (J) es, de hecho, equivalente a la limitación superior de la integral

$$\int \ln \frac{f^{\Theta}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < \infty, \tag{2}$$

donde $f^{\Theta}(x) = \sup_{t\in\Theta} f_t(x)$.

Demostración del teorema 2. El hecho de que de $(A_\delta^{\hat{\Delta}})$ resulte $(A_\delta^{\hat{\delta}})$ y (J) es evidente. Ahora supongamos que se cumplen $(A_\delta^{\hat{\delta}})$ y (J). Si admitimos que $((A_\delta^{\hat{\Delta}})$ no tiene lugar, existirán sucesiones $t_k \to t \in \Theta$, $\Delta_k \to 0$, $\varepsilon_k \to 0$ tales, que

$$\int \ln \frac{f_{t_k}^{\Delta}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) > -\varepsilon_k.$$

Aquí la función subintegral es mayorada, en virtud de la condición (J), por la función superiormente integrable, por eso, en virtud del lema de Fatou,

$$\lim_{k\to\infty} \sup \int \ln \frac{f_{t_k}^{\Delta_k}(x)}{f_\theta(x)} f_\theta(x)\mu(dx) \leqslant \int \ln \frac{f_t^{\circ}(x)}{f_\theta(x)} f_\theta(x)\mu(dx) < 0.$$

Hemos obtenido la contradicción que demuestra el teorema. ◁

Ahora expondremos unas condiciones bastante más simples, que demuestran el cumplimiento de $(A_\delta^{\hat{\delta}})$ y (J) y, por lo tanto, la conciliabilidad fuerte de la e.v.m.

Definición 1. Diremos que $f_t(x)$ pertenece a la clase D_0, si para cada $t \in \Theta$ existe un conjunto $C_t \in \mathfrak{B}_{\mathscr{X}}$, $\mathbf{P}_\theta(C_t) = 1$ en el que $f_t(x)$ es continua respecto a t: $f_{t_k}(x) \to f_t(x)$ cuando $t_k \to t$, $x \in C_t$.

Además de las $f_t(x)$ continuas (respecto a t) en el conjunto C, $\mathbf{P}_\theta(C) = 1$ independiente de t, a la clase D_0 también pertenecen, por supuesto, otras funciones, tales, por ejemplo, para las cuales $f_t(x)$ en el plano (t, x) tiene líneas de discontinuidad aisladas y desprovistas de partes paralelas al eje x. Así será, en particular, si $f_t(x)$, como función de x, tiene discontinuidades aisladas en los puntos $x_t^{(1)}$, $x_t^{(2)}$, ..., que dependen continuamente de t.

Teorema 3. *Si $f_t(x) \in D_0$ y se cumple (J), entonces también se cumple la condición $(A_\delta^{\hat{\delta}})$ y, por lo tanto, la e.v.m. $\hat{\theta}^*$ es fuertemente conciliable.*

Demostración. Si $f_t(x) \in D_0$, entonces $f_t^{\circ}(x) = f_t(x)$ cuando $x \in C_t$ y, por lo tanto,

$$\int \ln \frac{f_t^{\circ}(x)}{f_\theta(x)} f_\theta(x)\mu(dx) = -\varrho(\theta, t) < 0. \quad \triangleleft$$

Corolario 1. *Si* $f_t(x) \in D_0$ *está limitada, y la integral*

$$\int f_\theta(x) \ln f_\theta(x) \mu(dx) \tag{3}$$

es finita, la e.v.m. es fuertemente conciliable.

La afirmación del corolario 1 se deduce directamente del teorema 3, ya que el carácter limitado de $f_t(x)$ y la finitud de la integral (3) conducen a (J).

Corolario 2. *Si*

$$\varphi(\Delta) = \int \sup_{|u|<\Delta} |f_{t+u}(x) - f_t(x)| \mu(dx) \to 0 \tag{4}$$

cuando $\Delta \to 0$, *la e.v.m. es fuertemente conciliable.*

Demostración. Hagamos uso del teorema 3. La pertenencia de $f_t(x) \in D_0$ es evidente, ya que (4) puede cumplirse tan sólo en el caso en que $f_{t+u}(x) \to f_t(x)$ cuando $u \to 0$ para c.t. $[\mu]$ valores de x.

Luego,

$$\int f_t^\Delta(x) \mu(dx) \leqslant \varphi(\Delta) + \int f_t(x) \mu(dx) = \varphi(\Delta) + 1,$$

y la condición (4) también significa la integrabilidad de $f_t^\Delta(x)$. Como $\ln \dfrac{f_t^\Delta(x)}{f_\theta(x)} \leqslant \dfrac{f_t^\Delta(x)}{f_\theta(x)} - 1$, de aquí obtenemos que la integral en las condiciones (J) no supera

$$\int f_t^\Delta(x) \mu(dx) - 1 \leqslant \varphi(\Delta). \quad \triangleleft$$

En vez de (4) podríamos exigir la convergencia a cero de la magnitud

$$\varphi_1(\Delta) = \int \sup_{|u|\leqslant\Delta} (\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)})^2 \mu(dx),$$

ya que $\varphi(\Delta)$ se puede estimar con ayuda de $\varphi_1(\Delta)$ utilizando la desigualdad

$$\varphi(\Delta) \leqslant \int \sup_{|u|\leqslant\Delta} |\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)}| \sup_{|u|\leqslant\Delta} |\sqrt{f_{t+u}(x)} + \sqrt{f_t(x)}| \mu(dx) \leqslant$$

$$\leqslant \varphi_1^{1/2}(\Delta) \left[\int \sup_{|u|\leqslant\Delta} (\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)} + 2\sqrt{f_t(x)})^2 \mu(dx) \right]^{1/2} \leqslant$$

$$\leqslant [2\varphi_1(\Delta)(\varphi_1(\Delta) + 4]^{1/2}.$$

Corolario 3. *Si* $f_t(x)$ *es derivable respecto a* t *para c.t.* $[\mu]$ *valores de* x, *y*

$$\int |\dot{f}_t(x)| \mu(dx) < c < \infty, \tag{5}$$

entonces la e.v.m. $\hat{\theta}^*$ *es fuertemente conciliable. La condición* (5) *siempre se cumple si la información de Fisher* $I(t)$ *está limitada.*

Aquí hemos llegado al mismo resultado que podríamos obtener del teorema 23.2. El método de demostración de este último (véanse los §§ 21, 23) muestra que el carácter limitado de $I(t)$ o (5) no son esenciales para la afirmación del corolario 3 si la distancia de Hellinger $\varrho_3(\mathbf{P}_\theta, \mathbf{P}_{\theta+\Delta})$ está uniformemente separada del cero cuando $|\Delta| \geqslant \delta > 0$.

Demostración. La pertenencia de $f_t(x) \in D_0$ es evidente. Para el cumplimiento de la condición (J) es suficiente, como hemos visto en la demostración del corolario 2, la integrabilidad de $f_t^{\Delta}(x)$. Pero

$$\int f_t^{\Delta}(x)\mu(dx) \leqslant \int \left[f_t(x) + \int_{-\Delta}^{\Delta} |f'_{t+u}(x)|du \right] \mu(dx) =$$

$$= 1 + \int_{-\Delta}^{\Delta} \left[\int |f'_{t+u}(x)|\mu(dx) \right] du \leqslant 1 + 2\Delta c.$$

Queda hacer uso del teorema 3. La última afirmación del corolario 3 se deduce de la desigualdad de Cauchy — Buniakovski, ya que, en virtud de esta desigualdad, $\int |f_t'(x)|\mu(dx) \leqslant I^{1/2}(t)$. ◁

Corolario 4. *Sea θ el parámetro de desplazamiento de la familia $f_\theta(x) = f(x-\theta)$, $\int f(x) \ln f(x)dx > -\infty$. Si la función $f(x)$ está limitada (de lo contrario el método de verosimilitud máxima pierde su sentido (véase el § 26)) y tiene un conjunto B de puntos de discontinuidad, cuya medida de Lebesgue de clausura $\mu(B^c)$ es igual a cero, entonces la e.v.m. $\hat{\theta}^*$ es fuertemente conciliable.*

Demostración. Verifiquemos el cumplimiento de las condiciones del teorema 3. La condición (J) se cumple de modo evidente. La pertenencia de $f_t(x) \in D_0$ se desprende de la definición de D_0 en que es necesario poner $C_t = \overline{B^c} - t$ (este es el desplazamiento del conjunto $\overline{B^c}$ en t, y $\overline{B^c}$ es la adición a la clausura del conjunto B). En vista de que el conjunto $\overline{B^c}$ está abierto, $x - t \in \overline{B^c} - t$ conduce a $x - t_k \in \overline{B^c} - t$ para $|t_k - t|$ bastante pequeñas. Esto quiere decir que $f(x - t_k) \to f(x - t)$. El corolario queda demostrado.

Cabe señalar que en las condiciones del corolario 4 es inútil suponer que se ha cumplido la condición (A_0), puesto que ésta se cumple automáticamente. Si admitamos que (A_0^δ) no tiene lugar, llegaremos a la periodicidad de la función $f(x)$, lo que es imposible.

En cuanto a las condiciones del corolario 4, señalaremos que la condición de "continuidad" de $f(x)$, enunciada en este corolario, es muy débil. Pero, por lo visto, tampoco esta condición es esencial. Lo confirma, en cierta medida, el ejemplo siguiente.

Ejemplo 1. Sea $f(x)$ una función arbitraria que tiene un portador limitado
$(a, b) = \{x: f(x) > 0\}$. Entonces

$$\mathbf{P}_\theta(|\hat{\theta}^* - \theta| > \delta) \leqslant (1 - F_0(a + \delta))^n + F_0^n(b - \delta), \tag{6}$$

donde $F_\theta(x) = \int_{-\infty}^{x} f_\theta(y)dy$. La desigualdad (6) significa la conciliabilidad fuerte de $\hat{\theta}^*$. Esto se deduce de las relaciones que tienen la forma siguiente:

$$\{\hat{\theta}^* - \theta > \delta\} \subset \left\{ \prod_{l=1}^{n} f_{\theta+\delta}(x_l) > 0 \right\} \subset \bigcap_{l=1}^{n} \{x_l \geqslant a + \theta + \delta\},$$

$$\mathbf{P}_\theta(\hat{\theta}^* - \theta > \delta) \leqslant [1 - F_\theta(a + \theta + \delta)]^n = [1 - F_0(a + \delta)]^n.$$

Desde cierto punto de vista la condición de finitud de la integral $\int f(x) \ln f(x)\, dx$ en el corolario 4 tampoco es esencial: se puede construir fácilmente un ejemplo cuando esta integral se convierte en $-\infty$ y la condición (J) queda cumplida.

De las observaciones del § 2.18 se desprende que todo lo dicho en el corolario 4 y después de éste conserva por completo su validez para el parámetro de escala.

§ 28. Resultados de los §§ 23—27 para el caso del parámetro multidimensional

En este párrafo trasladaremos al caso multidimensional todos los resultados principales de los §§ 23—27. Dichos resultados serán expuestos en el mismo orden que en los párrafos indicados, con la particularidad de que sólo nos detendremos en los momentos donde el carácter multidimensional modifica la formulación del resultado o exige la modificación de los razonamientos.

Así pues, supongamos $\theta \in \Theta \subset R^k$, $k > 1$. Las enunciaciones de las condiciones (A_μ), (A_c) y (A_0), al igual que las definiciones de la relación de verosimilitud

$$Z(u) = \frac{f_{\theta+u}(X)}{f_\theta(X)}$$

y la distancia de Hellinger

$$r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta) = \int (\sqrt{f_{\theta+u}(x)} - \sqrt{f_\theta(x)})^2 \mu(dx),$$

no están relacionadas de ningún modo con la dimensión.

1. Desigualdades para la relación de verosimilitud (resultados del § 23). Para estudiar el comportamiento de la función $Z(u)$ en el entorno del cero

necesitaremos la condición siguiente: *la función* $\sqrt{f_\theta(x)}$ *es derivable respecto a* θ, *y la matriz de información de Fisher*

$$I(\theta) = |I_{ij}(\theta)| = \left\| \mathbf{M}_\theta \frac{\partial}{\partial \theta_i} l(x_1, \theta) \frac{\partial}{\partial \theta_j} l(x_1, \theta) \right\|, \tag{1}$$

para todos $\theta \in \Theta$, *está limitada y definida positivamente.*

Dada esta condición, del teorema 21.3A resulta que para todos θ,

$$0 < g \leqslant \frac{r(u)}{|u|^2} \leqslant h = \frac{1}{4} \sup_\theta \operatorname{Sp} I(\theta) < \infty. \tag{2}$$

Aquí y en lo sucesivo $|u|$ significa la norma euclídea $|u| = \sqrt{u_1^2 + \ldots + u_k^2}$ del vector $u = (u_1, \ldots, u_k)$.

La primera afirmación del teorema 23.1 y su demostración se trasladan al caso multidimensional sin camios algunos, ya que, de hecho, las mismas no están relacionadas con la dimensión.

Teorema 1. *Si se cumple* (2), *entonces*

$$\mathbf{M}_\theta Z^{1/2}(u) \leqslant e^{-ng|u|^2/2}.$$

Para generalizar el teorema 23.2 necesitaremos una condición adicional que consiste en que

$$\gamma \equiv \sup_\theta \mathbf{M}_\theta |l'(x_1, \theta)|^s < \infty \tag{3}$$

con cierto $s > k$.

Teorema 2 (análogo del teorema 23.2). *Si se cumplen las condiciones* (2) y (3), *entonces, con todos* z, $n \geqslant 1$

$$\mathbf{P}_\theta \left(\sup_{|v| > u} Z\left(\frac{v}{\sqrt{n}}\right) > e^z \right) \leqslant c\gamma e^{-z} + e^{-z/2} e^{-\beta u^2}, \tag{4}$$

donde $c < \infty$, $\beta > 0$ sólo dependen de k, g y s.

Para demostrar esta afirmación, en el caso unidimensional hemos utilizado la posibilidad de estimar $\sup_{u \in (0,\, 1)} p(u)$ por los valores de $p(0)$ y $\int_0^1 |p'(u)|du$. En el caso multidimensional, tal enfoque choca con dificultades, puesto que el valor máximo de $p(u)$ en cierta región $D \subset R^k$, $k > 1$, no puede ser estimado, hablando en general, por los valores de $p(u_0)$, $u_0 \in D$, y la integral de $p'(u)$ ($p'(u) = \operatorname{grad} p(u)$), por una curva registrada cualquiera de D. Existen, por lo menos, dos vías para superar esta dificultad.

La primera es absolutamente análoga al enfoque unidimensional y consiste en utilizar la estimación que tiene la siguiente forma (en esta fórmula,

para simplificar la escritura, nos limitamos al caso bidimensional $k = 2$):

$$\sup_{u \in K_{0,1}} p(u) \leqslant |p(0)| + \int_0^1 \left| \frac{\partial p((0,\ u_2))}{\partial u_2} \right| du_2 + \int_0^1 \left| \frac{\partial p((u_1,\ 0))}{\partial u_1} \right| \times$$

$$\times\ du_1 + \int_0^1 \int_0^1 \left| \frac{\partial^2 p(u)}{\partial u_1 \partial u_2} \right| du_1\ du_2,$$

donde $u = (u_1,\ u_2)$, $K_{0,1}$ es el cubo unitario $K_{0,1} = \{u\colon\ 0 \leqslant u_j \leqslant 1\ j = 1,\ \ldots,\ k\}$. Sin embargo, para utilizar este enfoque debemos suponer que existen derivadas de k-ésimo orden de la función $l_\theta(x)$ $(f_\theta(x))$ (véase la definición de la función p en el párrafo 23) y saber apreciar los valores medios (que necesitamos) de las derivadas de la función p del l-ésimo orden, $l \leqslant k$.

La segunda vía es más conveniente, ya que utiliza la posibilidad de estimar $\sup_{u \in K_{0,1}} p(u)$ a través de los valores de $p(0)$ y

$$\int_{K_{0,1}} |p'(u)|^s du\ \ (p'(u) = \operatorname{grad} p(u),\ u = (u_1,\ \ldots,\ u_k))$$

con cierto $s > k$ (cuando $s = k$ la estimación es imposible). En este caso, sin duda, debemos disponer de las estimaciones para $\mathbf{M}_\theta |p'(u)|^s$ cuando $s > k$. La obtención de todas las estimaciones aquí necesarias presenta ciertas dificultades y requiere mucho espacio. Por eso la demostración del teorema 2 para el caso multidimensional se da en el Suplemento VII.

También debemos señalar que en el libro editado en ruso se utilizó otro método de demostración del teorema 2 (véanse las observaciones bibliográficas referentes al Suplemento VII).

Las demostraciones de las afirmaciones acerca de la conciliabilidad de la e.v.m. y acerca de las estimaciones para los momentos en el punto 2 del § 23, no están relacionadas con la dimensión. Las propias afirmaciones se conservarán en la forma siguiente.

Teorema 3 (análogo del teorema 23.3). *Si se cumplen las condiciones* (2) *y* (4), *entonces para cualesquiera* z, $n \geqslant 1$ *es válida* (23.6) *sustituyendo el número* $g/4$ *por* β (véase el teorema 2).

Las afirmaciones de los corolarios 23.1 y 23.2 conservan por completo su validez sustituyendo igualmente $g/4$ por β.

2. Propiedades asintóticas de la relación de verosimilitud (resultados del § 24).

En el caso multidimensional, por condiciones (RR) entenderemos el conjunto de condiciones siguientes:

1) *Condiciones* (A_0), (A_c), (R).

2) *Derivabilidad continua de segundo orden respecto a* θ *dentro de* Θ, *de la función* $l(x, t)$ *para c.t.* $[\mu]$ *valores de* x. *En este caso se supone que las derivadas*

$$l''_{ij}(x, t) = \frac{\partial^2 l(x, t)}{\partial t_i \partial t_j}$$

admiten la mayorante $l(x)$ *que no depende de* t: $|l''_{ij}(x, t)| \leqslant l(x)$, *para la cual*

$$\mathbf{M}_t l(\mathrm{x}_1) = \int l(x) f_t(x) \mu(dx)$$

*converge uniformemente**) *en* $t \in \Theta$.

3) *Además, supondremos, siempre que sea necesario, que se cumple la condición* (3).

Al igual que en el caso unidimensional, necesitaremos las dos propiedades siguientes que se deducen de (*RR*):

1) *Posibilidad de derivar dos veces respecto a* θ *bajo el signo integral en la igualdad*

$$\int f_\theta(x) \mu(dx) = 1,$$

que significa la validez de las relaciones

$$\int \frac{\partial}{\partial \theta_i} f_\theta(x) \mu(dx) = 0, \qquad \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(x) \mu(dx) = 0. \tag{5}$$

2) *Convergencia uniforme de la integral* $I(\theta)$:

$$\sup_\theta \mathbf{M}_\theta [(l'(\mathrm{x}_1, \theta))^2; \quad |l'(\mathrm{x}_1, \theta)| > N] \to 0 \tag{6}$$

cuando $N \to \infty$.

Estas propiedades se demuestran en el Suplemento VI. Para simplificar la exposición, las referidas propiedades pueden ser intoducidas en las condiciones (*RR*).

En virtud de las igualdades

$$l'_i(x, \theta) = \frac{1}{f_\theta(x)} \cdot \frac{\partial f_\theta(x)}{\partial \theta_i},$$

$$l''_{ij}(x, \theta) = \frac{1}{f_\theta(x)} \frac{\partial^2 f_\theta(x)}{\partial \theta_i \partial \theta_j} - \frac{1}{f_\theta^2(x)} \cdot \frac{\partial f_\theta(x)}{\partial \theta_i} \cdot \frac{\partial f_\theta(x)}{\partial \theta_j},$$

de las relaciones (5) resulta que

$$\mathbf{M}_\theta l'_i(\mathrm{x}_1, \theta) = 0,$$

$$\mathbf{M}_\theta l''_{ij}(\mathrm{x}_1, \theta) = -\mathbf{M}_\theta l'_i(\mathrm{x}_1, \theta) l'_j(\mathrm{x}_1, \theta) = -I_{ij}(\theta).$$

*) Véase la nota en la pág. 226, acerca de la convergencia uniforme en el § 24.

Al igual que en el caso unidimensional, las condiciones (RR) significan que tendrán lugar las afirmaciones de los teoremas del § 23 acerca de las estimaciones para

$$\sup_{|v|>v} Z(v/\sqrt{n}) \text{ y para } \sqrt{n}(\hat{\theta}^* - \theta).$$

Al cumplirse las condiciones (*RR*), también serán válidos los siguientes análogos de los lemas 24.1 y 24.2.

Lema 1. *Las funciones* $l''_{ij}(x, \theta)$ son continuas "*por término medio*":

$$\mathbf{M}_\theta \omega''_\Delta(x_1) \to 0$$

es uniforme respecto a θ *cuando* $\Delta \to 0$, *donde* $\omega''_\Delta(x) = \max\limits_{i,j} \sup\limits_{\theta, |u|<\Delta} |l''_{ij} \times$ $\times (x, \theta + u) - l''_{ij}(x, \theta)|$.

La demostración repite exactamente los razonamientos del lema 24.1. ◁

Pongamos

$$\gamma_n(\delta, \theta) = \sup_{\substack{\Delta \leqslant \delta \\ |\omega|=1}} \left| \frac{(L'(X, \theta + \omega\Delta), \omega) - (L'(X, \theta), \omega)}{n\Delta} + \omega I(\theta)\omega^T \right|.$$

Lema 2. (análogo del lema 24.2). *Supongamos que se cumplen las condiciones (RR) y que* $\delta_n > 0$ *es cualquier sucesión que converge a cero. Entonces, para* $X \in \mathbf{P}_\theta$

$$\gamma_n(\delta_n, \theta) \underset{\text{c.s.}}{\to} 0, \quad \gamma_n(\delta_n, \hat{\theta}^*) \underset{\text{c.s.}}{\to} 0.$$

En estas relaciones, los valores de $I(\theta)$ *e* $I(\hat{\theta}^*)$ *pueden sustituirse uno por otro.*

Demostración. Al igual que en el caso unidimensional, es suficiente convencerse de que $\gamma_n(\delta_n) \to 0$, donde

$$\gamma_n(\delta) = \sup_{\substack{\Delta \leqslant \delta \\ |\omega|=1}} \left| \frac{(L'(X, \theta + \omega\Delta), \omega) - (L'(X, \theta), u)}{n\Delta} - \frac{\omega L''(X, \theta)\omega^T}{n} \right|.$$

Pero $\gamma_n(\delta_n) \leqslant \frac{1}{n} \sum\limits_i \sum\limits_{k,j} \omega''_{\delta_n}(x_i)|\omega_k\omega_j|$, donde $\omega''_\delta(x)$ es el módulo máximo de continuidad de las funciones $l''_{ij}(x, \theta)$. Como

$$\sum_{k,j} |\omega_k\omega_j| \leqslant k|\omega|^2 = k,$$

entonces

$$\gamma_n(\delta_n) \leqslant \frac{k}{n} \sum_i \omega''_{\delta_n}(x_i). \tag{7}$$

La demostración ulterior se base en el lema 1 y repite exactamente los razonamientos del lema 24.2. ◁

La generalización del teorema 24.1 para el caso multidimensional aquí es el

Teorema 4. *Supongamos que se cumplen las condiciones (RR) y que* $\delta_n > 0$, $n = 1, 2, \ldots$, *es cualquier sucesión convergente a cero. En este caso, si* $X \in \mathbf{P}_\theta$, *para u tales, que* $|u/\sqrt{n}| \leqslant \delta_n$,

$$Y(u) \equiv \ln Z(u/\sqrt{n}) = (\xi_n, u) - \frac{1}{2} uI(\theta)u^T(1 + \varepsilon_n(X, \theta, u)), \tag{8}$$

donde $|\varepsilon_n(X, \theta, u)| \leqslant \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$, $\xi_n = \frac{1}{\sqrt{n}} \operatorname{grad} L(X, \theta) = \frac{1}{\sqrt{n}} L'(X, \theta) \in$ $\in \Phi_{0, I(\theta)}$.

El valor de $u^* = \sqrt{n}(\hat{\theta}^* - \theta)$ *con el que* $Y(u)$ *alcanza su valor máximo es representable en la forma*

$$u^* = \xi_n I^{-1}(\theta)(E + \varepsilon_n(X, \theta)), \quad \varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0, \tag{9}$$

donde E es la matriz unidad. Además,

$$2Y(u^*) = \xi_n I^{-1}(\theta)\xi_n^T(1 + \varepsilon_n(X, \theta)) \in$$

$$\in \frac{1}{2} \xi I^{-1}(\theta)\xi^T \in H_k, \quad \xi \in \Phi_{0, I(\theta)}. \tag{10}$$

A la par con (8) *es válida la representación*

$$Y(u) - Y(u^*) = \frac{1}{2} (u - u^*)I(\theta)(u - u^*)^T(1 + \varepsilon_n(X, \theta, u)),$$

$$|\varepsilon_n(X, \theta, u)| \leqslant \varepsilon_n(X, \theta).$$

En todas las afirmaciones mencionadas se puede sustituir $I(\theta)$ *por* $I(\hat{\theta}^*)$.

Al igual que en el § 24, en este párrafo, por $\varepsilon_n(X, \theta)$ entendemos las distintas sucesiones que poseen la propiedad de $\varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$ respecto a $\mathbf{P}_\theta$.

También debemos señalar que el miembro principal en (8) puede ser escrito de la forma siguiente:

$$\xi_n u^T - \frac{1}{2} uI(\theta)u^T =$$

$$= -\frac{1}{2} (u - \xi_n I^{-1}(\theta))I(\theta)(u - \xi_n I^{-1}(\theta))^T + \frac{1}{2} \xi_n I^{-1}(\theta)\xi_n^T.$$

Esto corresponde a la densidad de una distribución normal multidimensional con media $\xi_n I^{-1}(\theta)$ y con matriz de segundos momentos $I^{-1}(\theta)$.

La demostración del teorema 4 es completamente análoga a la del teorema 24.1. Del lema 2, cuando $\Delta \leqslant \delta_n$, obtenemos

$$(L'(X, \theta + \Delta\omega), \omega) = (L'(X, \theta), \omega) - \\ - n\Delta\omega I(\theta)\omega^T(1 + \varepsilon_n(X, \theta, \Delta\omega)), \quad |\varepsilon_n(X, \theta, \Delta\omega)| \leqslant \varepsilon_n(X, \theta).$$

Integrando esta igualdad respecto a Δ de 0 a $|u|/\sqrt{n}$ y poniendo $\omega = u/|u|$, obtenemos

$$L(X, \theta + u/\sqrt{n}) - L(X, \theta) = \int_0^{|u|/\sqrt{n}} (L'(X, \theta + \Delta u), u)d\Delta =$$

$$= \frac{|u|}{\sqrt{n}} (L'(X, \theta), \omega) - \frac{|u|^2}{2} \omega I(\theta)\omega^T(1 + \varepsilon_n(X, \theta, u)) =$$

$$= (\xi_n, u) - \frac{1}{2} uI(\theta)u^T(1 + \varepsilon_n(X, \theta, u)), \quad |\varepsilon_n(X, \theta, u)| \leqslant \varepsilon_n(X, \theta).$$

Aquí, según el teorema central multidimensional de límite (véase el suplemento V),

$$\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(x_i, \theta) \in \Phi_{0, I(\theta)}.$$

La representación (8) queda demostrada. Las demás afirmaciones del teorema se demuestran absolutamente igual que en el teorema 24.1, teniendo en cuenta tan sólo las modificaciones de mostradas relacionadas con la multidimensión. La relación

$$\frac{1}{2} \xi I^{-1}(\theta)\xi^T \in \mathbf{H}_k$$

en (10) se deduce de las propiedades de la distribución normal (véase el punto 4 del § 2.2). ◁

Con arreglo a la relación (10) también es útil la siguiente

Observación 1. La matriz $I^{-1}(\theta)$, junto con $I(\theta)$, es positivamente definida, y existe una matriz $I^{-1/2}(\theta)$ que es la raíz cuadrada de $I^{-1}(\theta)$, o sea, una matriz que satiface la relación

$$I^{-1/2}(\theta)I^{-1/2}(\theta) = I^{-1}(\theta).$$

En efecto, si cierta matriz $\mathbf{M} > 0$ (está positivamente definida), entonces existe una matriz ortogonal C para la cual $CMC^T = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$ es una matriz diagonal con elementos positivos $\lambda_i > 0$ en la diagonal. Si ponemos ahora $\mathbf{M}^{1/2} = C^T \operatorname{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})C$, obtenemos, evidentemente, la raíz cuadrada de $\mathbf{M}$.

Valiéndonos de esto y de la simetría de la matriz $I^{-1}(\theta)$, podemos (10) escribir en la forma

$$\frac{1}{2} (\xi_n I^{-1/2}(\theta))(\xi_n I^{-1/2}(\theta))^T.$$

Aquí el vector $\eta_n = \xi_n I^{-1/2}(\theta)$ es, evidentemente, la suma normalizada de los vectores aleatorios igualmente distribuidos, con una media nula y una matriz de segundos momentos

$$\mathbf{M}_\theta(\xi_n I^{-1/2}(\theta))^T(\xi_n I^{-1/2}(\theta)) = \mathbf{M}_\theta I^{-1/2}(\theta)\xi_n^T\xi_n I^{-1/2}(\theta) = E,$$

puesto que

$$\mathbf{M}_\theta \xi_n^T \xi_n = \mathbf{M}_\theta(l'(x_1, \theta))^T(l'(x_1, \theta)) = I(\theta).$$

Esto significa que según el teorema central multidimensional del límite, $\xi_n I^{-1/2}(\theta) \in \Phi_{0,E}$.

Teorema 5 (análogo del teorema 24.2). *Supongamos que se cumplen las condiciones del teorema* 24.2 *para* $\theta \in R^k$ *multidimensional y para* $\alpha = \beta/2$ (β *está definido en el teorema* 2). *En este caso*

$$J \equiv \int w(u^* - u)q(\theta + u/\sqrt{n})Z(u/\sqrt{n})\Pi(du) = e^{Y(u^*)}q(\theta) \times$$

$$\times \left[\int w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)I(\theta)(u - u^*)^T\right\}\Pi(du) + \varepsilon_n(X, \theta)\right]. \tag{11}$$

Si Π *es la medida de Lebesgue, y* $\Pi(du) = du$, *entonces*

$$J = \frac{(2\pi)^{k/2}}{\sqrt{|I(\theta)|}} e^{Y(u^*)}q(\theta)(\mathbf{M}w(\eta) + \varepsilon_n(X, \theta)), \tag{12}$$

donde $\varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$, $\eta \in \Phi_{0, I^{-1}(\theta)}$ (*la sucesión* $\varepsilon_n(X, \theta)$ *es vectorial si* $w(t)$ *es una función vectorial*).

El teorema 5 se demuestra igual que el teorema 24.2, puesto que la demostración de este último no está relacionada con la dimensión.

3. Propiedades de la e.v.m. (resultados del § 25). Aquí siempre supondremos que se cumplen las condiciones (RR).

El análogo del teorema 25.1 tendrá la forma siguiente.

Teorema 6. *La e.v.m.* $\hat{\theta}^*$ *es una estimación asintóticamente normal, con la particularidad de que la convergencia*

$$u^* = (\hat{\theta}^* - \theta)\sqrt{n} \in \Phi_{0, I^{-1}(\theta)}$$

tiene lugar junto con los momentos de cualquier orden. En particular,

$$\mathbf{M}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) \to I^{-1}(\theta). \tag{13}$$

Además, para cualquier función continua $w(t)$ *tal, que* $|w(t)| < e^{\beta|t|^2/2}$ (*el número* β *está definido en el teorema* 2),

$$\mathbf{M}_\theta w(u^*) \to \mathbf{M}w(\eta), \quad \eta \in \Phi_{0, I^{-1}(\theta)}.$$

La relación (13) significa que $\hat{\theta}^* \in K_{\Phi,2}$.

La afirmación del teorema 6 se desprende del teorema 4 (véase (9)) y del análogo multidimensional de corolario 23.2 que se deduce del teorema 3 (compárese con la demostración del teorema 25.1). ◁

Definamos la clase $\tilde{K}_0$ como población de las estimaciones θ^* para las cuales el desplazamiento $b(\theta) = (b_1(\theta), ..., b_k(\theta)) = \mathbf{M}_\theta\theta^* - \theta$ posee las propiedades

$$|b(\theta)| = o(1/\sqrt{n}), \quad b_{ij}(\theta) = \frac{\partial b_i(\theta)}{\partial \theta_j} \to 0$$

cuando $n \to \infty$.

El análogo de los teoremas 25.2 y 25.3 aquí tiene la misma forma.

Teorema 7. *$\hat{\theta}^*$ es una estimación asintóticamente R-eficiente. Además, $\hat{\theta}^* \in \tilde{K}_0$ también es asintóticamente eficiente en $\tilde{K}_0$.*

El carácter asintóticamente R-eficiente de $\hat{\theta}^*$, equivalente a (13), tiene lugar evidentemente. La pertenencia de $\hat{\theta}^* \in \tilde{K}_0$ y la eficacia asintótica en $\tilde{K}_0$ se demuestran completamente igual que en el caso unidimensional.

Pasemos ahora a examinar la propiedad del carácter asintóticamente bayesiano. El carácter asintóticamente R-bayesiano de la estimación θ^* significa, por definición, que (compárese con el § 20)

$$\mathbf{M}(\theta^* - \theta)^T(\theta^* - \theta) = J/n + o(1/n), \quad J = \int I^{-1}(t)\mathbf{Q}(dt). \tag{14}$$

El carácter asintóticamente bayesiano de θ^* significa

$$\lim_{n \to \infty} \sup [nv(\theta^*) - nv(\theta_Q^*)] \leqslant 0, \tag{15}$$

donde θ_Q^* es la estimación bayesiana que minimiza $v(\theta^*) = \mathbf{M}(\theta^* - \theta) \times$ $\times V(\theta^* - \theta)^T$ para cualquier matriz V definida no negativamente.

Teorema 8 (análogo del teorema 25.4). *$\hat{\theta}^*$ es una estimación asintóticamente R-bayesiana. Si la distribución a priori $\mathbf{Q}$ tiene densidad respecto a la medida de Lebesgue en Θ, entonces $\hat{\theta}^*$ es una estimación asintóticamente bayesiana.*

La demostración es completamente análoga a la del teorema 25.4. La relación (14) para $\theta^* = \hat{\theta}^*$ se deduce del hecho de que

$$\lim_{n \to \infty} \mathbf{M}n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) =$$

$$= \mathbf{M} \lim_{n \to \infty} \mathbf{M}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) = \mathbf{M}I^{-1}(\theta) = J.$$

El paso límite bajo el signo de la esperanza matemática (o sea, de la integral) aquí es legítimo, ya que la magnitud $\mathbf{M}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta)$ está limitada por una constante que no depende ni de n ni de θ (compárese con el corolario 23.2).

Para demostrar (15) notemos que, conforme al § 20, la desigualdad integral de Rao—Cramer, en el caso cuando $\mathbf{Q}$ tiene densidad, reviste el aspecto

$$\mathbf{M}n(\theta^* - \theta)^T(\theta^* - \theta) \geqslant J + o(1).$$

Esto significa que

$$nv(\theta_Q^*) \geqslant \sum v_{ij}J_{ij} + o(1),$$

donde $|J_{ij}| = J$, $|v_{ij}| = V$. Por otro lado, en virtud de (14) cuando $\theta^* = \hat{\theta}^*$,

$$nv(\hat{\theta}^*) = \sum v_{ij}J_{ij} + o(1).$$

Es evidente que de estas relaciones se deduce (15) cuando $\theta^* = \hat{\theta}^*$. ◁

Los análogos de los teoremas 25.5 y 25.6 también tendrán lugar. Por ejemplo, del teorema 5 se desprende

Teorema 9 (análogo del teorema 25.6). *Supongamos que* $X \in \mathbf{P}_\theta$ *y que* θ *es un punto interior arbitrario de* Θ. *Si* $q(t)$ *es la densidad arbitraria continua y positiva (dentro de* Θ*) de la distribución a priori, entonces*

$$\sqrt{n}(\hat{\theta}^* - \theta_Q^*) \underset{\text{c.s.}}{\to} 0,$$

donde θ_Q^* *es la estimación bayesiana correspondiente a* $q(t)$.

El carácter asintóticamente minimax de $\hat{\theta}^*$ puede ser establecido igualmente que en el teorema 25.7, con ayuda del análogo multidimensional del criterio asintóticamente minimax en el corolario 20.3:

$$\lim_{n\to\infty} \sup_{t\in\Gamma} \mathbf{M}_t n(\hat{\theta}^* - \theta)V(\hat{\theta}^* - \theta)^T = \sup_{t\in\Gamma} \sum I_{ij}^{-1}(\theta)v_{ij}$$

$$|I_{ij}^{-1}(\theta)| = I^{-1}(\theta),$$

y con ayuda del carácter uniforme de convergencia en (13), la cual se deducirá de los resultados del párrafo siguiente.

En el caso del parámetro multidimensional $\hat{\theta}^*$, cuando su dimensión k es grande, las propiedades de la optimalidad asintótica de θ deben utilizarse con cuidado. Es necesario observar que la relación n/k sea grande (el número de observaciones para un parámetro escalar). De lo contrario las deducciones pueden resultar erróneas.

Ejemplo 1. En el laboratorio se comprueba la concentración de n soluciones. Cada una de las n concentraciones desconocidas $\mu_1, \ldots, \mu_n$ se verifica dos veces. Se supone que la varianza σ^2 de todas n observaciones $(x_1, y_1), \ldots, (x_n, y_n)$ es igual, y que las propias observaciones son independientes y están distribuidas normalmente, así que

$$f_\theta(X) = \frac{1}{\sigma^{2n}(2\pi)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum[(x_i - \mu_i)^2 + (y_i - \mu_i)^2]\right\},$$

donde

$$\theta = (\mu_1, \ldots, \mu_n, \sigma^2).$$

Las e.v.m. para μ_i aquí son iguales a

$$\hat{\mu}_i^* = \frac{1}{2}(x_i + y_i).$$

Es evidente que estas estimaciones no están desplazadas y no son conciliables. La e.v.m. para σ^2 es igual a

$$(\hat{\sigma}^2)^* = \frac{1}{4n}\sum (x_i - y_i)^2 \xrightarrow[P]{} \sigma^2/2 \text{ cuando } n \to \infty.$$

Esta estimación proporciona con gran fiabilidad un valor falso para el parámetro σ^2 (dos veces menor).

4. Cálculo aproximado de la e.v.m. El contenido de § 26 conservará por completo su validez en el caso multidimensional si por $[L''(X, t)]^{-1}$ entendemos la matriz inversa a $L''(X, t)$.

5. Propiedades de la e.v.m. al faltar las condiciones de regularidad (resultados de § 27). Las condiciones de conciliabilidad de θ, enunciadas en los teoremas 27.1—27.3, de hecho no están relacionadas con la dimensión. La demostración de estos teoremas se conserva por completo con una exactitud de hasta las modificaciones evidentes debidas al hecho de que el conjunto Θ ahora ha de ser recubierto (en virtud de la condición (A_c)) no por un número finito de intervalos, sino por un número finito de esferas. También se puede decir lo mismo en cuanto a los corolarios 27.1—27.4.

§ 29. Uniformidad respecto a θ, de las propiedades asintóticas de la relación de verosimilitud y de las estimaciones de verosimilitud máxima

En las investigaciones posteriores, principalmente en los §§ 13—15 de capítulo siguiente, serán útiles las afirmaciones de los §§ 24, 25 y 28 en su aspecto uniforme en cuanto a θ. La mayoría de estas afirmaciones (digamos, las que tratan de la $\mathbf{P}_\theta$-distribución límite de $(\hat{\theta}^* - \theta)\sqrt{n}$) han sido obtenidas suponiendo que θ es un punto registrado de Θ. Ahora nos interesará qué sucederá si θ no ha sido registrado y cambia junto con n. Está claro que en este caso junto con n también cambiarán las distribuciones $\mathbf{P}_\theta$, así que cada muestra X_n tendrá su "propia" distribución para $n = 1, 2, \ldots$

Llegamos, pues, al esquema de series (véase [11]), para el cual las enunciaciones de los principales teoremas del límite serán algo diferentes. En particular, la ley fuerte de los grandes números pierde, hablando en general, su sentido, ya que las variables aleatorias sujetas a investigación dejan de ser dadas (para diferentes n) en un espacio probabilístico.

1. **Ley uniforme de los grandes números y teorema central del límite.** Sea $X \in \mathbf{P}_\theta$, $\eta_{n,\theta} = \eta_n(X, \theta)$.

Definición 1. *Diremos que la sucesión $\eta_{n,\theta}$ converge uniformemente en probabilidad hacia la constante $a(\theta)$, si para cualquier $\varepsilon > 0$, cuando $n \to \infty$,*

$$\sup_{\theta \in \Theta} \mathbf{P}_\theta(|\eta_{n,\theta} - a(\theta)| > \varepsilon) \to 0.$$

Esta relación se escribirá en la forma "$\eta_{n,\theta} \underset{P_\theta}{\to} a(\theta)$ *uniformemente respecto a* θ".

Definición 2. Diremos que $\eta_{n,\theta}$ *converge en distribución hacia la variable aleatoria* η_θ *uniformemente respecto a* θ si para cualquier función continua y limitada φ, cuando $n \to \infty$,

$$\sup_{\theta} |\mathbf{M}_{\theta\varphi}(\eta_{n,\theta}) - \mathbf{M}_\varphi(\eta_\theta)| \to 0. \tag{1}$$

Esta relación es escribirá en la forma "$\eta_{n,\theta} \Rightarrow \eta_\theta$ *uniformemente respecto a* θ". Ese mismo sentido le conferiremos a la relación "$\eta_{n,\theta} \in \mathbf{G}_\theta$ uniformemente respecto a θ", donde $\mathbf{G}_\theta$ significa la distribución η_θ.

Le proponemos al lector que él mismo compruebe el hecho de que si las funciones de distribución η_θ son continuas uniformemente respecto a θ, la relación (1) es equivalente a

$$\sup_{\theta,\ x} |\mathbf{P}_\theta(\eta_{n,\theta} < x) - \mathbf{P}(\eta_\theta < x)| \to 0.$$

Nótese que la convergencia uniforme $\eta_{n,\theta} \underset{P_\theta}{\to} a(\theta)$ y la convergencia uniforme en distribución $\eta_{n,\theta} \Rightarrow a(\theta)$ hacia la variable aleatoria degenerada $a(\theta)$ son equivalentes.

Nótese también que para la convergencia uniforme conservarán su validez los principales teoremas de continuidad. Por ejemplo, si H es una función continua, de la convergencia uniforme $\eta_{n,\theta} \Rightarrow \eta_\theta$ se deduce la convergencia uniforme

$$H(\eta_{n,\theta}) \Rightarrow H(\eta_\theta). \tag{2}$$

Estas afirmaciones se deducen directamente de las definiciones.

En el Suplemento V hemos demostrado los siguientes teoremas "uniformes" del límite.

Supongamos que $X \in \mathbf{P}_\theta$ y que $a(x, \theta)$ es una función vectorial medible dada : $\mathscr{X} \times \Theta \to R^l$. Examinemos las sumas

$$s_n(\theta) = \sum a(x_i, \theta)$$

de los vectores aleatorios independientes que dependen del parámetro $\theta \in \Theta$

tanto directamente a través de la función $a(x, \theta)$, como también a través de la distribución de $x_i \in \mathbf{P}_\theta$.

Recordemos que la integral $\int \psi(x, \theta)\mathbf{P}_\theta(dx)$ se llama convergente uniformemente respecto a θ en la región Θ, si

$$\sup_{\theta \in \Theta} \int_{|\psi(x, \theta)| > N} |\psi(x, \theta)|\mathbf{P}_\theta(dx) \to 0$$

cuando $N \to \infty$.

Teorema 1 (ley uniforme de los grandes números). *Si la integral* $a(\theta) = \int a(x, \theta)\mathbf{P}_\theta(dx)$ *converge uniformemente respecto a* $\theta \in \Theta$, *entonces, cuando* $n \to \infty$,

$$\frac{s_n(\theta)}{n} \underset{P_\theta}{\to} a(\theta)$$

uniformemente respecto a θ.

Corolario 1. *Si la sucesión* $\{\theta_n\} \in \Theta$, *entonces en las condiciones del teorema* 1,

$$\mathbf{P}_{\theta_n}\left(\left|\frac{s_n(\theta_n)}{n} - a(\theta_n)\right| > \varepsilon\right) \to 0.$$

Este hecho será designado

$$\frac{s_n(\theta_n)}{n} - a(\theta_n) \underset{\mathbf{P}_{\theta_n}}{\to} 0.$$

Al examinar el teorema central del límite, para las sumas $s_n(\theta)$ será más cómodo suponer $a(\theta) = 0$. (Esto no es la limitación de la generalidad, ya que podemos examinar nuevos sumandos $a^1(x_i, \theta) = a(x_i, \theta) - a(\theta)$). Pongamos $a^2(\theta) = \mathbf{M}_\theta(a^T(x_1, \theta)a(x_1, \theta))$ y designemos por $a_j(x_1, \theta)$, $j = 1, 2, \ldots, l$ las coordenadas de los vectores $a(x_1, \theta)$.

Teorema 2 (teorema central uniforme del límite). *Supongamos que las integrales* $\int (a_j^2(x, \theta))\mathbf{P}_\theta(dx)$, $j = 1, \ldots, l$ *convergen uniformemente en* Θ. *Entonces*

$$\eta_{n,\theta} = \frac{s_n(\theta)}{\sqrt{n}} \Rightarrow \eta_\theta \in \Phi_{0, a^2(\theta)}$$

uniformemente respecto a θ.

2. Variantes uniformes de los teoremas de las propiedades asintóticas de la relación de verosimilitud y de las estimaciones de verosimilitud máxima. Nótese previamente que, al cumplirse las condiciones (RR), los resultados del § 23 serán uniformes respecto a θ por su propia forma, ya que los segundos miembros de las desigualdades en los teoremas 23.1 — 23.3 (y en los teoremas 28.1 — 28.3) no dependen de θ.

Pasemos a los resultados de los §§ 24 y 28 acerca del comportamiento asintótico de $Z(u/\sqrt{n})$.

Las afirmaciones de los lemas 24.1, 28.1, 24.2 y 28.2 pueden hacerse uniformes respecto a θ.

Lema 1. *Cuando* $\Delta \to 0$

$$\sup_{\theta} \mathbf{M}_\theta \omega''_\Delta(x_1) \to 0, \tag{3}$$

donde $\omega_\Delta(x_1)$ *es el módulo máximo de continuidad de las funciones* $l''_{ij}(x, \theta)$.

Demostración. La validez de (3) para un θ registrado ha sido demostrada en el lema 28.1. Si en este caso admitimos la ausencia de uniformidad respecto a θ, llegaremos al hecho de que existen $\varepsilon > 0$ y sucesiones $\theta_n \to \theta \in \Theta$, $\Delta_n \to 0$ tales, que

$$\mathbf{M}_{\theta_n} \omega''_{\Delta_n}(x_1) > \varepsilon. \tag{4}$$

Suponiendo, para abreviar, $\omega''_{\Delta_n}(x_1) = \omega''$, obtenemos

$$\mathbf{M}_{\theta_n}\omega'' = \mathbf{M}_{\theta_n}(\omega''; f_{\theta_n}(x_1) \leqslant 2f_\theta(x_1)) + \mathbf{M}_{\theta_n}(\omega''; f_{\theta_n}(x_1) > 2f_\theta(x_1), l(x_1) \leqslant N) + \mathbf{M}_{\theta_n}(\omega''; f_{\theta_n}(x_1) > 2f_\theta(x_1), l(x_1) > N).$$

Aquí el primer sumando no excede $2\mathbf{M}_\theta\omega''$ y converge a cero en virtud del lema 28.1. El segundo sumando no supera $2NJ_n$, donde

$$J_n = \int_{f_{\theta_n}(x) > 2f_\theta(x)} f_{\theta_n}(x)\mu(dx) = 1 - \int_{f_{\theta_n}(x) \leqslant 2f_\theta(x)} f_{\theta_n}(x)\mu(dx) \to 0$$

según el teorema de la convergencia mayorada. Por fin, el último sumando no supera $\mathbf{M}_{\theta_n}(2l(x_1); l(x_1) > N)$ y, en virtud de (RR), puede hacerse, escogiendo N, tan pequeño como se quiera. Hemos obtenido la contradicción con (4), lo cual demuestra el lema.

Lema 2. *La afirmación del lema* 28.2 *se conservará si la convergencia casi segura en ella se sustituye por la convergencia* $\gamma_n(\delta_n, \theta) \underset{P_\theta}{\to} 0$, $\gamma_n(\delta_n, \hat{\theta}^*) \underset{P_\theta}{\to} 0$ uniforme respecto a θ.

Demostración. Seguiremos la demostración del lema 28.2. Señalemos previamente que, en virtud del teorema 1 y de la convergencia uniforme de la integral en (RR),

$$L''(X, \theta)/n \underset{P_\theta}{\to} -I(\theta)$$

uniformemente respecto a θ (la convergencia de las matrices se entiende por elementos). Además, de los teoremas 23.3 y 28.3 se deduce que $\hat{\theta}^* \underset{P_\theta}{\to} \theta$ uniformemente respecto a θ. De aquí se desprende que en la relación $\gamma_n(\delta_n, \theta) \underset{P_\theta}{\to} 0$ (véase el lema 28.2) podemos sustituir $I(\theta)$ por $L''(\theta)/n$ y por $I(\hat{\theta}^*)$.

En virtud de la desigualdad (28.7), el problema de estimación de $\gamma_n(\delta_n, \theta)$ se reduce a la estimación de

$$\bar{\omega}''_{\delta_n}(X) = \frac{1}{n}\sum_{i=1}^{n} \omega''_{\delta_n}(x_i, \theta),$$

donde $\omega''_\Delta(x, \theta)$ es el módulo máximo de continuidad de las funciones $l''_{ij}(x, \theta)$. De la desigualdad de Chébishev obtenemos

$$\sup_\theta \mathbf{P}_\theta(\bar{\omega}''_{\delta_n}(X) > \varepsilon) \leqslant \frac{1}{\varepsilon}\sup_\theta \mathbf{M}_\theta\, \omega''_{\delta_n}(x_1, \theta).$$

Pero en virtud del lema 1, $\sup_\theta \mathbf{M}_\theta \omega''_\Delta(x_1, \theta) \to 0$ cuando $\Delta \to 0$. Esto demuestra que

$$\bar{\omega}''_{\delta_n}(X) \underset{P_\theta}{\to} 0, \quad \gamma_n(\delta_n, \theta) \underset{P_\theta}{\to} 0 \tag{5}$$

uniformemente respecto a θ.

Luego, de las desigualdades (24.6) resulta que el problema de estimación de $\gamma_n(\delta_n, \hat{\theta}^*)$ se reduce a la estimación de $\bar{\omega}''_{\delta_n+|\hat{\theta}^*-\theta|}(X)$. Como $\hat{\theta}^* - \theta \underset{P_\theta}{\to} 0$ uniformemente respecto a θ, de (5) obtenemos que

$$\bar{\omega}''_{\delta_n+|\hat{\theta}^*-\theta|}(X) \underset{P_\theta}{\to} 0, \quad \gamma_n(\delta_n, \hat{\theta}^*) \underset{P_\theta}{\to} 0$$

uniformemente respecto a θ. ◁

Teorema 3 (análogo del teorema 28.4). *Al cumplirse las condiciones (RR), las afirmaciones del teorema* 28.4 *se conservarán en las modificaciones siguientes:* $\varepsilon_n(X, \theta) \underset{P_\theta}{\to} 0$ *uniformemente respecto a* θ, $\xi_n \in \Phi_{0, I(\theta)}$, $2Y(u^*) \in H_k$ *uniformemente respecto a* θ.

La demostración del teorema se basa por completo en el lema 2, así como la demostración del teorema 28.4 se basa en el lema 28.2. Por eso la demostración requerida se obtiene mediante la introducción de modificaciones evidentes en la demostración del teorema 28.4, relacionadas con la sustitución (que resulta del lema 28.2) de la convergencia $\varepsilon_n(X, \theta) \underset{\text{c.s.}}{\to} 0$ por la convergencia uniforme $\varepsilon_n(X, \theta) \underset{P_\theta}{\to} 0$. Además, hay que añadir que

$$\xi_n = \frac{1}{n}\sum_{i=1}^{n} l(x_i, \theta) \in \Phi_{0, I(\theta)}$$

uniformemente respecto a θ, en virtud del teorema 2 y de la convergencia uniforme (28.6) de la integral $I(\theta)$ (ésta es la matriz de segundos momentos

para $l'(x_1, \theta)$), la cual se desprende de las condiciones (RR) (véase el Suplemento VI). De aquí y de las observaciones referentes a (2) obtenemos la convergencia uniforme

$$2Y(u^*) \in \mathbf{H}_k. \quad \triangleleft$$

Las mismas modificaciones que en el teorema 3 (en comparación con el teorema 28.4) pueden ser introducidas en los teoremas 28.5 y 28.6.

Citemos aquí los dos siguientes corolarios del teorema 3.

Teorema 4.

$$u^* = \sqrt{n}(\hat{\theta}^* - \theta) \in \Phi_{0, I^{-1}(\theta)} \tag{6}$$

uniformemente respecto a θ. *En este caso, para cualquier función* $w(x)$ *continua casi por doquier respecto a la medida de Lebesgue y tal que* $|w(x)| < Ce^{\beta|x|^{2/2}}$ (*el valor de* $\beta > 0$ *ha sido definido en el teorema* 28.2), *se cumple*

$$\sup_{\theta} |\mathbf{M}_\theta w(u^*) - \mathbf{M}w(\eta_\theta)| \to 0, \tag{7}$$

donde $\eta_\theta \in \Phi_{0, I^{-1}(\theta)}$.

Demostración. La primera afirmación se deduce de las relaciones

$$u^* = \xi_n I^{-1}(\theta)(E + \varepsilon_n(X, \theta)),$$
$$|\varepsilon_n(X, \theta)| \underset{P_\theta}{\to} 0, \quad \xi_n \in \Phi_{0, I(\theta)},$$

uniformes respecto a θ y contenidas en el teorema 3.

Para demostrar la segunda afirmación admitamos que (7) no es cierta. Entonces habrá $\delta > 0$ y sucesiones $\theta_n \to \theta \in \Theta$ tales, que

$$|\mathbf{M}_{\theta_n} w(u^*) - \mathbf{M}w(\eta_{\theta_n})| > \delta \tag{8}$$

para todos n.

Pero $\Phi_{0, I^{-1}(\theta_n)} \Rightarrow \Phi_{0, I^{-1}(\theta)}$ y, por consiguiente, en virtud de (6), la $\mathbf{P}_{\theta_n}$-distribución $u^*(w(u^*))$ converge débilmente a la distribución $\eta_\theta(w(\eta_\theta))$. Además, según el corolario 23.2 (véase también el § 28),

$$\sup_{\theta} \mathbf{M}_\theta w^{3/2}(u^*) \leqslant \sup_{\theta} \mathbf{M}_\theta \exp\{3(u^*)^2\beta/4\} < c_1 < \infty.$$

De aquí y de los teoremas de continuidad para los momentos se deduce que

$$\mathbf{M}_{\theta_n} w(u^*) \to \mathbf{M}w(\eta_\theta).$$

En vista de que $\mathbf{M}w(\eta_{\theta_n}) \to \mathbf{M}w(\eta_\theta)$, la relación obtenida contradice (8). $\triangleleft$

Sea $A_n \subset \mathscr{X}^n$.

Teorema 5. *Si* $\mathbf{P}_\theta(A_n) \to 0$, *entonces para cualquier* N *registrado,*

$$\sup_{|u| \leqslant N} \mathbf{P}_{\theta + u/\sqrt{n}}(A_n) \to 0.$$

Esta propiedad de las sucesiones de las distribuciones $\mathbf{P}_{\theta + u/\sqrt{n}}$ cuando $n \to \infty$ se llama *contigualidad* (véase [81]). La utilizaremos en el capítulo 3.

Demostración. Tenemos

$$\mathbf{P}_{\theta + u/\sqrt{n}}(A_n) = \mathbf{M}_\theta\{Z(u/\sqrt{n});\ A_n\} \leqslant$$
$$\leqslant \mathbf{M}_\theta(Z(u/\sqrt{n});\ A_n \cap \{Y(u) \leqslant c\}) + \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c) \leqslant$$
$$\leqslant e^c \mathbf{P}_\theta(A_n) + \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c).$$

Como $\mathbf{P}_\theta(A_n) \to 0$, para demostrar el teorema debemos examinar sólo $\sup_{|u| \leqslant N} \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c)$. Según el teorema 3,

$$Y(u) = (\xi_n,\ u) - \frac{1}{2} uI(\theta)u^T(1 + \varepsilon_n(X,\ \theta + u/\sqrt{n})) \in \Phi_{-\frac{1}{2}\sigma^2, \sigma^2} \qquad (9)$$

uniformemente respecto a u, donde $\sigma^2 = uI(\theta)u^T \leqslant N^2 \Lambda_k(\theta)$ cuando $|u| \leqslant N$, y $\Lambda_k(\theta)$ es el número máximo propio de la matriz $I(\theta)$. Como $\Phi_{-\frac{1}{2}\sigma^2, \sigma^2}((c,\ \infty)) \leqslant \Phi_{0, \sigma^2}((c,\ \infty))$, entonces, en virtud de la uniformidad en (9),

$$\lim_{n \to \infty} \sup_{|u| \leqslant N} \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c) \leqslant \sup_{|u| \leqslant N} \Phi_{0, \sigma^2}((c,\ \infty)) = \Phi_{0, N^2 \Lambda_k(\theta)}((c,\ \infty)).$$

Eligiendo c, este valor puede hacerse tan pequeño como se quiera. ◁

3. Algunos corolarios.

1) En el § 25 hemos enunciado el teorema 25.3 en el que se afirma, en particular, que $\hat{\theta}^* \in K^\circ$, donde K° es la clase de estimaciones asintóticamente centrales, la cual es definida por la relación (se examina el caso unidimensional)

$$\mathbf{P}_\theta(\hat{\theta}^* > \theta) \to 1/2$$

uniformemente respecto a θ. Del teorema 4 se deduce que la parte mencionada del teorema 25.3 es cierta, así que

$$\mathbf{P}_\theta(\hat{\theta}^* > \theta) = \mathbf{P}_\theta(\sqrt{n}(\hat{\theta}^* - \theta)I^{-1/2}(\theta) > 0) \to \Phi_{0,1}((0,\ \infty)) = 1/2$$

uniformemente respecto a θ. ◁

2) En el § 25 hemos enunciado el teorema 25.7 acerca del carácter asintóticamente minimax de $\hat{\theta}^*$. Para **demostrar** este teorema sólo queda establecer la validez del lema 25.1 de que

$$\lim_{n \to \infty} \sup_{\theta \in \Gamma} \mathbf{M}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta \in \Gamma} I^{-1}(\theta), \qquad (10)$$

donde Γ es cualquier segmento de Θ. Pero esta afirmación es el corolario directo de la convergencia de $\mathbf{M}_\theta n(\hat{\theta}^* - \theta)^2 \to I^{-1}(\theta)$, uniforme respecto a $\theta \in \Theta$, la cual hace válido el paso límite bajo el signo $\sup_{\theta\in\Gamma}$:

$$\lim_{n\to\infty} \sup_{\theta\in\Gamma} \mathbf{M}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} \lim_{n\to\infty} \mathbf{M}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} I^{-1}(\theta). \quad \triangleleft$$

La afirmación, que es análoga a (10) y asegura el carácter asintóticamente minimax de $\hat{\theta}^*$, tendrá lugar, evidentemente, también en el caso multidimensional:

$$\lim_{n\to\infty} \sup_{\theta\in\Gamma} \mathbf{M}_\theta n(\hat{\theta}^* - \theta) V(\hat{\theta}^* - \theta)^T = \sup_{\theta\in\Gamma} \sum v_{ij} I_{ij}^{-1}(\theta),$$

$$\|I_{ij}^{-1}(\theta)\| = I^{-1}(\theta)$$

para cualquier matriz V.

§ 30°. Acerca de los problemas estadísticos relacionados con las muestras de volumen aleatorio. Estimación sucesiva

El hecho de que las muestras de volumen aleatorio surgen en la práctica y son naturales, es confirmado por el ejemplo 18.3. Otro ejemplo está relacionado con la llamada estimación sucesiva (o progresiva), que se emplea en los casos cuando podemos realizar observaciones sucesivas, es decir, una tras otra, y cuando estamos interesados en minimizar el número de tales observaciones, digamos, debido a su alto precio. En este caso, además de la regla de estimación (o sea, de construcción de la estimación θ^*) debemos establecer la regla de interrupción del experimento. Estas reglas pueden ser diferentes: por ejemplo, podemos sumar los precios dados $c(x_i)$ de las observaciones x_i hasta agotarse cierta cantidad admisible t. En este caso el momento ν de interrupción (número de la última observación o volumen de la muestra) será determinado como

$$\nu = \min \left\{k: \sum_{i=1}^{k} c(x_i) \geqslant t\right\},$$

esto es "el tiempo del primer del nivel t" en errar con saltos $c(x_i)$ (véase [11], capítulo 8). Se pueden sumar las "informaciones" $I(x_i, \theta) = (I'(x_i, \theta))^2$ e interrumpir las observaciones cuando sea alcanzado otra vez cierto nivel dado, etc.

En estos ejemplos ν es un momento markoviano, o sea, $\{\nu > n\} \in \sigma(x_1, \ldots, x_n)$, que constituye una de las suposiciones principales al examinar los problemas de estimación sucesiva. Al hacer tal suposición y al cumplirse varias condiciones adicionales menos esenciales, la desigualdad

de Rao — Cramer será conservada en la forma siguiente:

$$\mathbf{D}_\theta \theta^* \geqslant \frac{1}{I(\theta)\mathbf{M}\nu},$$

donde $\theta^* = \theta^*(x_1, \ldots, x_\nu)$ es la estimación no desplazada de θ, $I(\theta)$, es decir, la información de Fisher. La demostración de esta desigualdad es análoga a las demostraciones del § 16, para calcular la información de Fisher, contenida en la muestra $(x_1, \ldots, x_\nu)$, sólo se necesita utilizar la identidad de Wald (véase [11]).

Si ν depende de cierto parámetro t, como ocurrió en el ejemplo 18.3, así que $\nu \to \infty$ casi siempre cuando $t \to \infty$, entonces es posible construir las estimaciones asintóticamente óptimas con un error estándar asintóticamente equivalente a $(I(\theta)\mathbf{M}\nu)^{-1}$.

§ 31. Estimación por intervalo

1. Definiciones. Hasta ahora hemos estudiado las propiedades y los métodos de búsqueda de las mejores estimaciones *puntuales* de un parámetro desconocido que determina la distribución $\mathbf{P}_\theta$ de la familia $\mathscr{P} = \{\mathbf{P}_\theta\}$, correspondiente a la muestra X. Las estimaciones puntuales se utilizan en los casos cuando debemos llamar cierto número θ^* destinado al uso en vez de θ desconocido.

No obstante, también tiene gran aplicación otro enfoque de la cuestión.

Consideraremos θ como parámetro escalar (el caso multidimensional será examinado en el punto 6). Como sabemos, no es posible determinar exactamente θ basándose en una muestra dada. Pero podríamos tratar de indicar tal intervalo (θ^-, θ^+), el cual, con una probabilidad dada bastante alta, sea capaz de recubrir el valor desconocido de θ. En este caso es indudable que cuanto más estrecho sea este intervalo tanto mejor será. En muchos problemas se exige de antemano, digamos, aumentando el volumen de la muestra, construir tal intervalo (θ^-, θ^+) cuya anchura no supere las dimensiones dadas.

Definición 1. Supongamos que para $\varepsilon > 0$ dado existen variables aleatorias $\theta^* = \theta^*(\varepsilon, X)$ tales que

$$\mathbf{P}_\theta(\theta^-(\varepsilon, X) < \theta, \quad \theta^+(\varepsilon, X) > \theta) \geqslant 1 - \varepsilon. \tag{1}$$

Entonces el intervalo (θ^-, θ^+) se llama *intervalo confidencial para* θ *de nivel* $1 - \varepsilon$.

Es evidente que (1) se puede escribir en la forma

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) \geqslant 1 - \varepsilon.$$

El suceso que aquí está bajo el signo de probabilidad, consiste en que el *intervalo aleatorio* (θ^-, θ^+) *ha cubierto el valor desconocido de* θ. Leer

este suceso como "θ toma un valor perteneciente al intervalo (θ^-, θ^+)" sería un poco menos exacto, ya que θ, hablando en general, no es aleatorio.

Los valores de θ^* se denominan *fronteras de los intervalos confidenciales*, y el número $1 - \varepsilon$, *coeficiente o nivel de confianza.*

Por lo tanto, la diferencia entre la estimación por intervalo y la estimación puntual consiste en lo siguiente.

1) El intervalo confidencial como estimación es "menos exacto", ya que se señala un conjunto entero de posibles valores de θ.

2) Por otro lado, la afirmación "$\theta \in (\theta^-, \theta^+)$ con probabilidad $\geqslant 1 - \varepsilon$" es real, mientras que el suceso $\theta = \theta^*$ tiene, por lo general, una probabilidad igual a cero.

En calidad de ε suele escogerse un número pequeño. Basándose en éste, se construyen $\theta^*(\varepsilon, X)$ y luego, basándose en la muestra, se declara que $\theta \in (\theta^-(\varepsilon, X), \theta^+(\varepsilon, X))$. Procediendo de este modo nos equivocaremos en una larga serie de experimentos, aproximadamente en el $100\varepsilon\%$ de todos los casos. Por ejemplo, si $\varepsilon = 0{,}001$, el error puede ocurrir una vez en 1000 casos, aproximadamente.

Declarando justa la relación $\theta \in (\theta^-, \theta^+)$, utilizamos el hecho de que si cierto suceso tiene la probabilidad ε y este ε es pequeño, entonces prácticamente es imposible que tal suceso se produzca durante un solo experimento. Un pasajero, tomando el avión cree intuitivamente en ello con seguridad. Le basta saber que la probabilidad de que el vuelo se termine felizmente es bastante alta (a pesar de que conoce que esta probabilidad no es igual a 1). Precisamente tal enfoque es la base para construir muchos procedimientos estadísticos.

Destaquemos primeramente un caso, cuando la construcción de los intervalos confidenciales es sobre todo natural y puede ser realizada sin grandes dificultades. Es el llamado caso bayesiano que ya hemos examinado en los §§ 10, 11 y 20.

2. Construcción de intervalos confidenciales en el caso bayesiano. Aquí supondremos que el parámetro θ se escoge *aleatoriamente*, con una densidad a priori conocida de distribución $q(t)$ respecto a cierta medida λ en Θ. Luego se realiza la muestra $X \in \mathbf{P}_\theta$ y necesitamos construir el intervalo confidencial para el valor elegido de θ.

Si se cumple la condición (A_μ), en este caso, como sabemos del § 10, existe una distribución a posteriori de θ (convencional respecto a X) que tiene una densidad de

$$q(t/X) = \frac{f_t(X)q(t)}{\int f_u(X)q(u)\lambda(du)}$$

respecto a la medida λ. Esto quiere decir que en calidad de $\theta^*(\varepsilon, X)$ es

suficiente tomar dos números cualesquiera $\theta^{\pm}$, para los cuales

$$\int_{\theta^-}^{\theta^+} q(u/X)\lambda(du) = 1 - \varepsilon$$

(o bien $\geqslant 1 - \varepsilon$ si $\int_{-\infty}^{t} q(u/X)\lambda(du)$ cambia al variar t discretamente). En otros términos, en calidad de θ^- y θ^+ conviene tomar las cuantilas de distribución a posteriori que tienen los órdenes $1 - \varepsilon_2$ y ε_1, respectivamente, para todos ε_1 y ε_2, tales que $\varepsilon_1 + \varepsilon_2 = \varepsilon$.

Aquí, a distinción del caso no bayesiano, en la relación $\theta^- \leqslant \theta \leqslant \theta^+$ son aleatorios todos los tres elementos : las fronteras del intervalo de θ^* y la propia magnitud θ.

No es difícil ver que en el procedimiento descrito existe cierta arbitrariedad relacionada con la elección de los números ε_1 y ε_2. A veces esta arbitrariedad es eliminada por el propio planteamiento del problema, por ejemplo, cuando nos es importante establecer únicamente la frontera confidencial superior o inferior. En este caso conviene poner igual a 0 uno de los números ε_1, ε_2 y hacer infinita la frontera respectiva. Sin embargo, si las fronteras desempeñan un papel simétrico, es natural escoger ε_i de modo que el intervalo (θ^-, θ^+) se haga más corto en la medida de lo posible. Para las distribuciones $q(t/X)$ próximas a las distribuciones simétricas, esto se alcanza cuando $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$.

3. Construcción de intervalos confidenciales en el caso general. Intervalos confidenciales asintóticos. Los principales métodos de construcción de intervalos confidenciales se basan en la utilización de estimaciones puntuales. Examinemos al principio el enfoque asintótico de la construcción de intervalos confidenciales.

Definición 2. Supongamos que $X = [X_\infty]_n \in \mathbf{P}_\theta$ y que para $\varepsilon > 0$ establecido existen variables aleatorias $\theta^*(\varepsilon, X)$ tales que

$$\lim_{n\to\infty} \inf \mathbf{P}_\theta(\theta^-(\varepsilon, X) < \theta < \theta^+(\varepsilon, X)) \geqslant 1 - \varepsilon. \tag{2}$$

En este caso el intervalo (θ^-, θ^+) se llama intervalo *asintótico confidencial de nivel* $1 - \varepsilon$.

En esta definición es necesario subrayar que en realidad se trata de la sucesión de intervalos (θ_n^-, θ_n^+) determinados para cada n. Formalmente, el concepto de intervalo asintótico confidencial, con arreglo a una muestra de volumen registrado, es insustancial. No obstante, la relación (2) se utiliza con grandes n al igual que se utiliza el teorema central del límite para el cálculo aproximado de las distribuciones de las sumas de un número finito de variables aleatorias.

En los apartados precedentes hemos visto que la mayoría de las estimaciones puntuales examinadas eran asintóticamente normales. Más abajo se expone la construcción de los intervalos asintóticos confidenciales basados en tales estimaciones.

Sea θ^* la estimación asintóticamente normal:

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2(\theta)}, \tag{3}$$

y $\sigma(\theta)$ es una función continua. Como $\theta^* \to \theta$, la última condición significa que $\sigma(\theta^*) \xrightarrow[p]{} \sigma(\theta)$. De aquí y de (3), según el segundo teorema de continuidad, resulta que

$$\frac{(\theta^* - \theta)\sqrt{n}}{\sigma(\theta^*)} \in \Phi_{0,\,1}. \tag{4}$$

Designemos por λ_δ la cuantila de distribución normal de orden $1 - \delta$, o sea, un número tal que $\Phi_{0,1}((-\infty, \lambda_\delta)) = 1 - \delta$, o bien $\mathbf{P}(|\xi| < \lambda_\delta) = 1 - 2\delta$ si $\xi \in \Phi_{0,1}$. Al disponer de $\varepsilon > 0$ registrado, para $\lambda_{\varepsilon/2}$ introduzcamos temporalmente una designación más breve, suponiendo

$$\lambda_{\varepsilon/2} = \beta.$$

Entonces de (4) se deduce

$$\lim_{n\to\infty} \mathbf{P}_\theta \left(\left| \frac{(\theta^* - \theta)\sqrt{n}}{\sigma(\theta^*)} \right| < \beta \right) = 1 - \varepsilon.$$

Pero esta relación se puede escribir en la forma

$$\lim_{n\to\infty} \mathbf{P}_\theta(\theta^* - \beta\sigma(\theta^*)/\sqrt{n} < \theta < \theta^* + \beta\sigma(\theta^*)/\sqrt{n} = 1 - \varepsilon.$$

Ahora bien, los números

$$\theta^{\pm} = \theta^* \pm \beta\sigma(\theta^*)/\sqrt{n} \tag{5}$$

satisfacen la definición 2 y, por consiguiente, son las fronteras del intervalo asintótico confidencial de nivel $1 - \varepsilon$.

Si ahora, para la muestra X dada y registrada, de volumen n, construimos el intervalo (5), su nivel real se distinguirá, hablando en general, de ε, pero se distinguirá poco si n es bastante grande. Por eso los intervalos asintóticos confidenciales deben tratarse con cierto cuidado, aclarando previamente a partir de qué n la probabilidad del suceso $\{\theta \in (\theta^-, \theta^+)\}$ es con bastante exactitud aproximada por el valor límite. Por regla general, cuanto menor sea ε tanto mayor será la exigencia en cuanto al volumen de la muestra n. El volumen necesario también depende de la distribución $\mathbf{P}_\theta$ y de la estadística θ^*.

Ejemplo 1. Supongamos que $X \in \Gamma_{\alpha,1}$ y que utilizamos la estimación eficiente $\alpha^* = \frac{n-1}{n\bar{x}}$. En los ejemplos 4.1 y 16.1 hemos establecido que

$$\mathbf{M}_\alpha \alpha^* = \alpha, \quad \mathbf{D}_\alpha \alpha^* = \frac{\alpha^2}{n-2},$$

así que aquí $\sigma^2(\alpha) = \alpha^2$. La relación (5) nos da

$$\alpha^{\pm} = \frac{n-1}{n\bar{x}}(1 \pm \beta/\sqrt{n}). \tag{6}$$

¿A qué realmente es igual el nivel de este intervalo?

Necesitamos hallar $\Gamma_{\alpha,1}$, o sea, la probabilidad de la desigualdad

$$\frac{n-1}{n\bar{x}}(1-\beta/\sqrt{n}) < \alpha < \frac{n-1}{n\bar{x}}(1+\beta/\sqrt{n})$$

o bien, que es lo mismo, la probabilidad de la desigualdad

$$1 - \beta/\sqrt{n} < \frac{n\alpha\bar{x}}{n-1} < 1 + \beta/\sqrt{n},$$

donde $n\alpha\bar{x} \in \Gamma_{1,n}$. Como α es el parámetro de escala, entonces $2n\alpha\bar{x} \in \in \Gamma_{1/2,n} = \mathbf{H}_{2n}$. Así pues, el nivel exacto del intervalo (6) es igual a

$$\int_{2(n-1)(1-\beta/\sqrt{n})}^{2(n-1)(1+\beta/\sqrt{n})} \gamma_{1/2,n}(x)\,dx, \tag{7}$$

donde $\gamma_{1/2,n}$ está definido en el § 2*).

Cuando $\varepsilon = 0{,}05$ y $n = 30$, tenemos $\beta = 1{,}96$, $(n-1)(1-\beta/\sqrt{n})/n \approx$ $\approx 0{,}6201$, $(n-1)(1+\beta/\sqrt{n})/n \approx 1{,}3126$.

Ahora bien, el intervalo asintótico confidencial de nivel $1-\varepsilon = 0{,}95$ con arreglo al caso $n = 30$, es el intervalo $(0{,}620/\bar{x},\ 1{,}313/\bar{x})$.

Si hacemos uso de las tablas de distribución χ^2 con 60 grados de libertad, en virtud de (7) descubriremos que el nivel exacto de significación de este intervalo confidencial constituye (con una exactitud de hasta tres signos) $0{,}937 = 1 - 0{,}063$. En este caso los "aportes" de los extremos izquierdo y derecho del referido intervalo no son equivalentes ni mucho menos (compárese con la aproximación normal) y constituyen 0,010 y 0,053, respectivamente.

Para $n = 50$ el intervalo asintótico confidencial de nivel, igual a 0,95, tendrá la forma $(0{,}708/\bar{x},\ 1{,}252/\bar{x})$. El nivel real de su significación será

*) La observación de que $\Gamma_{1/2,n} = \mathbf{H}_{2n}$ es útil, ya que permite, para el cálculo de $\Gamma_{\alpha,\lambda}$ (si 2λ es entero), utilizar las tablas de la distribución χ^2 dadas en el suplemento, así como en muchos otros manuales de estadística matemática.

igual a 0,942 = 1 − 0,058 (los aportes equivalen a 0,014 y 0,044, respectivamente). Está claro que si continuamos aumentando n, dichos aportes se aproximarán con 0,025.

Volvamos a examinar el intervalo confidencial (5) que hemos construido con ayuda de la estimación asintóticamente normal θ^*. A distinción del caso bayesiano, aquí hay una arbitrariedad relacionada con la elección de la estimación θ^*. La forma de las fronteras del intervalo muestra que se pueden obtener las dimensiones dadas del intervalo, tanto aumentando el volumen de la muestra n (lo que por diferentes causas no siempre es realizable) como disminuyendo posiblemente $\sigma(\theta^*)$. Aquí llegamos a la conclusión importante de que siendo iguales los volúmenes de la muestra, la estimación de menor dispersión $\sigma(\theta)$ dará el mejor intervalo confidencial. Ahora bien, *los mejores intervalos asintóticos confidenciales se obtendrán al utilizar las estimaciones asintóticamente eficientes.*

Siempre que se cumplan las condiciones (RR) y que θ^* pertenezca a la clase $\bar{K}_0 \cap K_{\Phi,2}$ (véanse los §§ 8 y 16) el mejor intervalo asintótico confidencial tendrá las siguientes fronteras:

$$\theta^{**} = \theta^* \pm \beta/\sqrt{nI(\theta^*)},$$

donde θ^* es cualquier estimación asintóticamente eficiente, por ejemplo, la e.v.m.

Algunos otros métodos de construcción de intervalos asintóticos confidenciales se examinarán en el punto 6.

4. Construcción del intervalo confidencial exacto mediante una estadística dada. Supongamos que en calidad de estadística hemos escogido la estimación θ^*. Entonces, mediante esta estimación, sería natural buscar el intervalo confidencial simétrico de nivel $1 - \varepsilon$ en la forma $\theta^* \pm \Delta(\varepsilon, X)$ o en la forma $\theta^*(1 \pm \Delta(\varepsilon, X))$, así como se hizo en el ejemplo antes examinado. No obstante, si tratamos de realizar este plan, resultará que la cosa no es tan simple, ya que en el caso general las fronteras $\pm\Delta(\varepsilon, X)$ dependerán del parámetro desconocido θ: pues $\Delta(\varepsilon, X)$ debe ser elegido de la condición

$$\mathbf{P}_\theta(\theta^* - \Delta(\varepsilon, X) < \theta < \theta^* + \Delta(\varepsilon, X)) \geqslant 1 - \varepsilon,$$

donde θ aquí entra, de manera esencial y muy compleja, antes que nada a través de la propia distribución $\mathbf{P}_\theta$.

Por eso, para construir los intervalos confidenciales mediante una estimación dada θ^*, se necesita cierta estructura especial.

En la construcción expuesta más abajo, a la par con la estimación θ^* puede participar cualquier estadística S. Designemos con el símbolo $\mathbf{G}_\theta$ la distribución de S y pongamos $G_\theta(x) = \mathbf{G}_\theta((-\infty, x))$.

Definición 3. Diremos que la estadística S, *en cuanto a su distribución, depende monótonamente de* θ si para todos x, $\theta_1 < \theta_2$

$$\mathbf{G}_{\theta_1}((x, \infty)) \leqslant \mathbf{G}_{\theta_2}((x, \infty))$$

o bien, que es lo mismo,

$$G_{\theta_1}(x) \geqslant G_{\theta_2}(x). \tag{8}$$

Todas las estimaciones razonables θ^* suelen poseer esta propiedad.

Si la dependencia monótona $G_\theta(x)$ de θ es continua, entonces la ecuación

$$G_\theta(x) = \gamma$$

es siempre resoluble respecto a θ para cada $\gamma \in (0, 1)$. Designemos por $b(x, \gamma)$ la solución de esta ecuación.

Teorema 1. *Si* $\varepsilon_1 + \varepsilon_2 = \varepsilon$, *la estadística* S, *en cuanto a su distribución, depende monótonamente de* θ, *y la función* $G_\theta(x)$ *es continua respecto a* θ *y* x, *entonces los valores*

$$\theta^- = b(S, 1 - \varepsilon_2), \quad \theta^+ = b(S, \varepsilon_1)$$

formarán el intervalo confidencial de nivel $1 - \varepsilon$.

La demostración del teorema es casi evidente. Utilicemos el hecho de que si la función de distribución $F(x)$ es continua y $\xi \in F$, entonces $F(\xi) \in \mathbf{U}_{0,1}$ ($\mathbf{P}(F(\xi) < x) = \mathbf{P}(\xi < F^{-1}(x)) = F(F^{-1}(x)) \equiv x$). En virtud de esta observación, $G_\theta(S) \in \mathbf{U}_{0,1}$ y, por lo tanto,

$$\mathbf{P}_\theta(\varepsilon_1 < G_\theta(S) < 1 - \varepsilon_2) = 1 - \varepsilon,$$
$$\mathbf{P}_\theta(b(S, 1 - \varepsilon_2) < \theta < b(S, \varepsilon_1)) = 1 - \varepsilon. \quad \triangleleft$$

Con frecuencia es cómodo realizar en dos etapas la "inversión" de la función $G_\theta(S)$, utilizada en el teorema. Primeramente $G_\theta(x)$ se invierte respecto a x, o sea, se determinan las cuantilas $G_\theta^{-1}(\gamma)$ como soluciones de las ecuaciones $G_\theta(x) = \gamma$, y luego se resuelven, respecto a θ, las ecuaciones

$$G_\theta^{-1}(\varepsilon_1) = S, \quad G_\theta^{-1}(1 - \varepsilon_2) = S.$$

Tales soluciones siempre existirán, ya que, según los datos del teorema, $G_\theta^{-1}(\gamma)$ depende monótona y continuamente de θ.

En la fig. 3 se muestran las curvas $y = G_\theta^{-1}(\varepsilon_1)$ e $y = G_\theta^{-1}(1 - \varepsilon_2)$ que definen para cada θ el campo de valores y, cuya probabilidad de entrar en el mismo, para cierta estimación $S = \theta^*$, es igual a $1 - \varepsilon$. Como ya hemos señalado, el procedimiento de construcción del intervalo confidencial es la inversión de las funciones

$$y = G_\theta^{-1}(\varepsilon_1), \quad y = G_\theta^{-1}(1 - \varepsilon_1),$$

o sea, la determinación de los puntos de intersección de las curvas de nivel $y = S$ que les corresponden. Los puntos de intersección obtenidos dan precisamente el intervalo requerido (θ^-, θ^+).

Si la condición de continuidad de $G_\theta(x)$ no se cumple, lo cual tendrá lugar para variables aleatorias discretas S, entonces, en general, el procedimiento expuesto y la afirmación del teorema 1 conservarán su validez, con la única diferencia de que, al definir respectivamente las cuantilas $G_\theta^{-1}(\gamma)$,

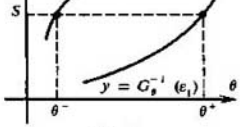

Fig. 3.

es necesario sátisfacer la desigualdad $\mathbf{G}_\theta((G_\theta^{-1}(\varepsilon_1), G_\theta^{-1}(1-\varepsilon_2))) \geqslant 1-\varepsilon$ en vez de la cual antes hemos tenido una igualdad exacta. En consonancia con esto, la afirmación del teorema 1 en este caso tendrá la forma

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) \geqslant 1-\varepsilon,$$

donde θ^* son las soluciones de las ecuaciones $G_\theta^{-1}(\varepsilon_1) = S$, $G_\theta^{-1}(1-\varepsilon_2) = S$. Además, llamaremos intervalo confidencial de nivel $1-\varepsilon$ el intervalo (θ^-, θ^+).

Si construimos el intervalo confidencial (θ^-, θ^+) con ayuda de la estimación θ^*, de la fig. 3 se deduce que este intervalo será tanto más estrecho cuanto más estrecho sea el intervalo $(G_\theta^{-1}(\varepsilon_1), G_\theta^{-1}(1-\varepsilon_2))$ o bien, que es lo mismo, cuanto más concentrada sea la distribución de θ^* cerca de θ. Ahora bien, aquí llegamos al mismo problema que en la teoría de las estimaciones puntuales, o sea, a la determinación de las estimaciones θ^* que aprecian θ de la forma más exacta.

El problema relacionado con la construcción de los mejores intervalos confidenciales se examina más detalladamente en el § 3.8.

En vista de que la forma de las funciones de distribución $G_\theta(x)$ suele ser bastante compleja incluso para las familias simples de distribuciones citadas en el § 2, el referido procedimiento de inversión de $G_\theta(x)$ en la práctica resulta muy difícil. Por eso el cálculo de las fronteras confidenciales está considerablemente tubulado. En el ejemplo siguiente, donde ilustramos la construcción de los intervalos confidenciales según el esquema descrito en el teorema 1, para simplificar la exposición utilizaremos la aproximación normal.

Ejemplo 2. Sea $X \in \mathbf{B}_p$. En calidad de estimación para p tomemos la estimación eficiente $p^* = \nu/n$, donde ν es el número de casos favorables en n pruebas (el número ν puede designar, por ejemplo, la cantidad de artículos desechados que han sido descubiertos durante la verificación de control de n muestras. Es necesario construir el intervalo confidencial para la porción de artículos defectuosos p).

Tenemos ($q = 1 - p$)

$$G_p(x) = \mathbf{P}_p(p^* < x) = \mathbf{P}_p\left(\frac{\nu - np}{\sqrt{npq}} < \frac{xn - np}{\sqrt{npq}}\right).$$

Conforme al teorema 1 debemos resolver la ecuación

$$G_p(p^*) = \gamma \tag{9}$$

para los valores γ iguales a $\varepsilon/2$ y $1 - \varepsilon/2$. Cuando n son grandes, en virtud del teorema central del límite, $G_p(x) \approx \Phi((x - p)n/\sqrt{npq})$, donde $\Phi(y) = \Phi_{0,1}((-\infty, y))$, y, por consiguiente, la ecuación (9) puede ser sustituida por su aproximación

$$\Phi((p^* - p)n/\sqrt{npq}) = \gamma, \quad \gamma = \varepsilon/2,\ 1 - \varepsilon/2,$$

o bien, que es lo mismo, $|(p^* - p)n/\sqrt{npq}| = \lambda_{\varepsilon/2} = \beta$,

$$(p^* - p)^2 = \beta^2 p(1 - p)/n.$$

Esta es la ecuación para las fronteras $p^{\pm}$ del intervalo confidencial, que no es otra cosa sino la ecuación de la elipse extendida para grandes n a lo largo de la bisectriz $p^* - p = 0$. Despejando p en esta ecuación, obtenemos

$$p^{\pm} \approx p^* \pm \beta\sqrt{p^*(1 - p^*)/n}.$$

No es difícil comprobar que obtendríamos ese mismo resultado si utilizáramos el enfoque asintótico expuesto en el punto 3.

Si n no es grande, conviene calcular $G_p(x)$ por la fórmula exacta

$$G_p(x) = \sum_{k < nx} C_n^k p^k (1 - p)^{n-k},$$

aplicando luego el procedimiento del teorema 1.

Supongamos, por ejemplo, que de $n = 10$ artículos $\nu = 2$ resultaron defectuosos. Entonces, cuando $\varepsilon = 0{,}05$, las fronteras exactas del intervalo confidencial son iguales a $p^- = 0{,}037$, $p^+ = 0{,}507$. La gran anchura del intervalo se explica por la poca información de que disponemos.

No obstante, si $n = 100$, $\nu = 20$, entonces, para $\varepsilon = 0{,}05$,

$$p^- = 0{,}137, \quad p^+ = 0{,}277.$$

Hemos tomado estas cifras de tablas especiales que dan la solución numérica del problema sobre los intervalos confidenciales para el número p, siendo diferentes n y ν (véase [8]).

5. Otros métodos de construcción de intervalos confidenciales.

En este apartado examinaremos ciertas generalizaciones del procedimiento antes propuesto, relacionado con la construcción de intervalos confidenciales.

Teorema 2. *Admitamos que en* $\Theta \times \mathscr{X}^n$ *existe una función* $G(\theta, x)$, *tal, que la distribución* $\mathbf{H}(B) = \mathbf{P}_\theta(G(\theta, X) \in B)$ *no depende de* θ. *Supongamos también, que* $G(\theta, x)$, *para cada* x, *es continua y monotóna respecto a* θ.

Admitamos luego, que y^-, y^+ *satisfacen la relación* $\mathbf{H}((y^-, y^+)) = 1 - \varepsilon$. *Entonces las estadísticas*

$$\theta^- = G^{-1}(y^-, X),\ \theta^+ = G^{-1}(y^+, X),\ \textit{si}\ G(\theta, \cdot)\uparrow,$$

y

$$\theta^- = G^{-1}(y^+, X),\ \theta^+ = G^{-1}(y^-, X),\ \textit{si}\ G(\theta, \cdot)\downarrow,$$

son las fronteras del intervalo confidencial de nivel $1 - \varepsilon$. *Aquí* $G^{-1}(y, X)$ *es la solución de la ecuación* $G(\theta, X) = y$.

Demostración. En virtud de la monotonía de $G(\theta, x)$ (supongamos, para precisar, que $G(\theta, x)$ crece respecto a θ), el suceso $\{G^{-1}(y^-, X) < \theta < G^{-1}(y^+, X)\}$ coincide con el suceso $A = \{y^- < G(\theta, X) < y^+\}$.

Por definición de $\mathbf{H}(\cdot)$ e $y^\pm$ tenemos

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) = \mathbf{P}_\theta(G^{-1}(y^-, X) < \theta < G^{-1}(y^+, X)) = \mathbf{P}_\theta(A) = \mathbf{H}((y^-, y^+)) = 1 - \varepsilon. \quad \triangleleft$$

Observación 1. En el teorema 1, en calidad de $G(\theta, X)$ hemos examinado la función $G_\theta(S)$. Además se ha cumplido $\mathbf{H} = \mathbf{U}_{0,1}$.

Observación 2. Se puede examinar el análogo asintótico del teorema 2, admitiendo la existencia de la sucesión de funciones $\{G_n(\theta, x)\}$ continuas y monótonas respecto a θ y tales que, cuando $n \to \infty$,

$$\mathbf{P}_\theta(G_n(\theta, X) \in B) \to \mathbf{H}(B),$$

donde $\mathbf{H}(\cdot)$ no depende de θ. Entonces obtendremos el método de construcción de intervalos asintóticos confidenciales, que generaliza el método de construcción de intervalos asintóticos confidenciales mediante estimaciones asintóticamente normales, expuesto en el punto 3.

Ahora proponemos un método más (a la par con el teorema 1) de elección de la función $G(\theta, x)$ que figura en el teorema 2.

Teorema 3. *Sea* $F_\theta(x) = \mathbf{P}_\theta(\mathbf{x}_1 < x)$, *con la particularidad de que*

1) $F_\theta(x)$ *es continua respecto a* x *para todos* $\theta \in \Theta$,

2) *$F_\theta(x)$ es continua y monótona respecto a θ para cualquier x registrado. Entonces la función*

$$G(\theta, x) = -\sum_{i=1}^{n} \ln (F_\theta(x_i))$$

satisface las condiciones del teorema 2.

Si los números y^* son tales que

$$\frac{1}{\Gamma(n)} \int_{y^-}^{y^+} x^{n-1} e^{-x} dx = 1 - \varepsilon, \tag{10}$$

entonces $\theta^* = G^{-1}(y^+, X)$ formarán las fronteras del intervalo confidencial de nivel $1 - \varepsilon$.

Demostración. Verifiquemos el cumplimiento de las condiciones del teorema 2. Como, según la condición 1), $F_\theta(x_i)$ distribuida uniformemente en $[0, 1]$, entonces $-\ln F_\theta(x_i) \in \Gamma_{1,1}$ y $G(\theta, X) \in \Gamma_{1,n}$. Con otras palabras, $\mathbf{P}_\theta(G(\theta, X) \in B) = \Gamma_{1,n}(B)$ y $\mathbf{H} = \Gamma_{1,n}$ no depende de θ. La monotonía y la continuidad de $G(\theta, x)$ se deducen, para cada x, de la condición 2). Además, en virtud de (10)

$$\mathbf{H}((y^-, y^+)) = \Gamma_{1,n}((y^-, y^+)) = 1 - \varepsilon. \quad \triangleleft$$

También se pueden señalar algunas otras construcciones de los intervalos confidenciales. En este caso, al igual que en la teoría de estimación puntual, en seguida surge la pregunta acerca de qué intervalo confidencial debe considerarse el mejor si se han obtenido varios intervalos. En el § 3.8 trataremos de los enfoques que existen en este caso. Sin embargo, de la exposición precedente está claro que, de hecho, el problema de búsqueda del intervalo confidencial óptimo es en mucho muy parecido al problema de estimación puntual óptima. También está claro que si construimos los intervalos confidenciales utilizando las estimaciones puntuales, conviene dar preferencia a los intervalos confidenciales construidos con ayuda de las mejores estimaciones.

La semejanza de los problemas de optimación de las estimaciones puntual y por intervalo puede ser ilustrada citando el ejemplo de la afirmación siguiente.

Teorema 4. *Examinemos el intervalo asintótico confidencial (θ^-, θ^+) de nivel $1 - \varepsilon$ y supongamos que la variable aleatoria $\theta^* = (\theta^+ + \theta^-)/2$ es la estimación asintóticamente normal y asintóticamente central* (véase el punto 2 del § 25), *y la magnitud $\Delta = (\theta^+ - \theta^-)/2$ es tal, que $\delta = \lim\inf_{n\to\infty} \sqrt{n}\Delta$ no depende de X. En este caso $\delta \geqslant \beta/\sqrt{I(\theta)}$.*

Esto quiere decir que la anchura del intervalo confidencial (θ^-, θ^+) no puede ser mucho menor que $2\beta/\sqrt{nI(\theta)}$, o sea, menor que la anchura del intervalo de nivel $1-\varepsilon$ construido con ayuda de la e.v.m. $\hat{\theta}^*$.

Demostración. Admitamos lo contrario. Entonces habrá una subsucesión de los números $\{n'\}$ para los cuales $\Delta\sqrt{n'} \to c\beta/\sqrt{I(\theta)}$, $c < 1$. Como $\theta^{\pm} = \theta^* \pm \Delta$, entonces

$$1-\varepsilon = \lim_{n'\to\infty} \mathbf{P}_\theta(\theta^- < \theta < \theta^+) = \lim_{n'\to\infty} \mathbf{P}_\theta(|\theta^* - \theta| < \Delta) =$$

$$= \lim_{n'\to\infty} \mathbf{P}_\theta(|\theta^* - \theta|\sqrt{n'} < c\beta/\sqrt{I(\theta)}) \leqslant \lim_{n\to\infty} \mathbf{P}_\theta(|\hat{\theta}^* - \theta|\sqrt{n} < c\beta/\sqrt{I(\theta)}). \quad (11)$$

La última desigualdad se deduce del hecho de que la e.v.m. $\hat{\theta}^*$ es asintóticamente eficiente en la clase $\bar{K}^\circ$ de estimaciones asintóticamente centrales (véase el teorema 25.4). En vista de que en (11) el segundo miembro es menor que $1-\varepsilon$, hemos obtenido la contradicción que demuestra el teorema. $\triangleleft$

6. Caso multidimensional. El concepto de intervalo confidencial se generaliza en el caso del parámetro multidimensional $\theta \in R^k$ en el concepto de región confidencial o de conjunto confidencial.

Definición 4. El subconjunto aleatorio*) $\Theta^* = \Theta^*(\varepsilon, X)$ del espacio paramétrico Θ se llama *conjunto confidencial de nivel* $1-\varepsilon$ si

$$\mathbf{P}_\theta(\Theta^* \ni \theta) \geqslant 1-\varepsilon. \quad (12)$$

Con otras palabras, el conjunto confidencial de nivel $1-\varepsilon$ recubre el valor real desconocido de θ con una probabilidad no menor de $1-\varepsilon$.

Definición 5. Si $X = [X_\infty]_n \in \mathbf{P}_\theta$, y si el conjunto aleatorio Θ^* satisface la relación

$$\lim_{n\to\infty} \inf \mathbf{P}_\theta(\Theta^* \ni \theta) \geqslant 1-\varepsilon,$$

entonces Θ^* se llama *conjunto asintótico confidencial de nivel* $1-\varepsilon$.

Los conjuntos confidenciales "exactos", incluso óptimos, se estudian en el § 8 del capítulo siguiente.

En lo que se refiere a los conjuntos asintóticos confidenciales, el principio de su construcción es el mismo de antes. Teniendo en cuenta el teorema 4, examinaremos a la vez los conjuntos confidenciales construidos con ayuda de la e.v.m. θ^*. Como sabemos, al cumplirse las condiciones (RR), $X \in \mathbf{P}_\theta$,

$$(\hat{\theta}^* - \theta)\sqrt{n}I^{1/2}(\theta) \in \Phi_{0,E}.$$

*) En este contexto diremos que el conjunto $\Theta^*(\varepsilon, X)$ es aleatorio si para cada t el conjunto $[X\colon t \in \Theta^*(\varepsilon, X))$ es medible y, por lo tanto, también diremos que la probabilidad (12) está definida (compárese con el § 3.8.).

De aquí se deduce que

$$n(\hat{\theta}^* - \theta)I(\theta)(\hat{\theta}^* - \theta)^T \in \mathbf{H}_k,$$
$$n(\hat{\theta}^* - \theta)I(\hat{\theta}^*)(\hat{\theta}^* - \theta)^T \in \mathbf{H}_k.$$

Con otras palabras, si h_ε significa la cuantila de orden $1 - \varepsilon$ de la distribución χ^2 con k grados de libertad, entonces

$$\lim_{n\to\infty} \mathbf{P}_\theta(n(\theta - \hat{\theta}^*)I(\hat{\theta}^*)(\theta - \hat{\theta}^*)^T < h_\varepsilon) = 1 - \varepsilon. \tag{13}$$

Hemos construido el conjunto asintótico confidencial Θ^* de nivel $1 - \varepsilon$ que es un elipsoide cuyo centro se encuentra en el punto $\hat{\theta}^*$ y cuyos ejes se definen por la matriz $nI(\hat{\theta}^*)/h_\varepsilon$. En este caso no es obligatorio calcular la matriz $I(\theta)$ para la construcción de Θ^*. Como sabemos, al cumplirse las condiciones (RR), $X \in \mathbf{P}_\theta$,

$$L(X, \theta) - L(X, \hat{\theta}^*) \approx -\frac{n}{2}(\theta - \hat{\theta}^*)I(\hat{\theta}^*)(\theta - \hat{\theta}^*)^T.$$

Por eso el elipsoide Θ^* definido en (13) puede representarse como la población de los valores de θ para los cuales

$$L(X, \theta) - L(X, \hat{\theta}^*) \geqslant -h_\varepsilon/2.$$

En el § 28 hemos determinado que el límite de la $\mathbf{P}_\theta$-probabilidad de esta desigualdad (véase la observación 28.2) es igual a $1 - \varepsilon$.

De aquí resulta, en particular, que en el caso unidimensional, las fronteras θ^* del intervalo asintótico confidencial de nivel $1 - \varepsilon$ pueden ser definidas como las soluciones de la ecuación

$$L(X, \theta) - L(X, \hat{\theta}^*) = -h_\varepsilon/2 = -\beta^2/2.$$

§ 32. Distribuciones muestrales exactas e intervalos confidenciales exactos para poblaciones normales

Entre todas las distribuciones citadas en el § 2, la distribución normal tiene la mayor aplicación. Por eso en este párrafo examinaremos especialmente la construcción de los intervalos confidenciales para los parámetros α y σ^2 de la distribución Φ_{α,σ^2}.

1. Distribuciones exactas de las estadísticas $\bar{x}$, S_0^2. Supongamos que $X \in \Phi_{0,1}$ y que $C = \|c_{ij}\|$ $(i, j = 1, 2, \ldots, n)$ es una matriz ortogonal.

Examinemos la distribución del vector n-dimensional $Y = XC$, $Y = (y_1, \ldots, y_n)$, $y_i = \sum_{j=1}^{n} x_j c_{ji}$.

Lema 1. *Si C es una matriz ortogonal, entonces $Y \in \Phi_{0,1}$, o sea, las coordenadas $y_1, \ldots, y_n$ son variables aleatorias independientes, $y_i \in \Phi_{0,1}$, $i = 1, 2, \ldots, n$.*

Demostración. Sea t un vector $(t_1, \ldots, t_n)$. La normalidad de la distribución de X significa que su función característica es igual a

$$\mathbf{M}e^{itX^T} = e^{-\frac{1}{2}tmt^T},$$

donde $m = \|m_{ij}\|$ es una matriz de segundos momentos, que en nuestro caso es igual a la matriz unidad E para la cual $tEt^T = \sum_{j=1}^{n} t_j^2$,

$$\mathbf{M}e^{itX^T} = e^{-\frac{1}{2}\sum_{j=1}^{n} t_j^2}.$$

La función característica de la distribución compatible $y_1, \ldots, y_n$ (o de la distribución del vector Y) tiene la forma

$$f(t) = \mathbf{M}e^{itY^T} = \mathbf{M}e^{itC^TX^T}.$$

Sustituyendo las variables $t = uC$ y notando que $CC^T = E$, obtenemos

$$f(t) = \mathbf{M}e^{iuCY^T} = \mathbf{M}e^{iuX^T} = e^{-\frac{1}{2}\sum_{j=1}^{n} u_j^2} = e^{-\frac{1}{2}\sum_{j=1}^{n} t_j^2}.$$

Esto quiere decir que Y tiene la misma función característica y, por lo tanto, la misma distribución que X. ◁

Ahora demostremos una afirmación llamada lema de Fisher, que es muy importante para la exposición ulterior.

Lema 2. *Supongamos, como antes, que* $X \in \Phi_{0,1}$, *que* C *es una matriz ortogonal y que* $Y = (y_1, \ldots, y_n) = XC$. *Entonces, la forma cuadrática*

$$T(X) = \sum_{i=1}^{n} x_i^2 - y_1^2 - \ldots - y_r^2$$

no depende de las variables aleatorias $y_1, \ldots, y_r$ *y tiene una distribución* χ^2 *con* $n - r$ *grados de libertad*:

La demostración es casi evidente, ya que después de aplicar la transformación ortogonal de C, obtenemos

$$T(X) = \sum_{i=1}^{n} y_i^2 - y_1^2 - \ldots - y_r^2 = y_{r+1}^2 + \ldots + y_n^2.$$

Solamente queda utilizar el lema 1. ◁

Pasemos ahora al estudio de la distribución compatible de las estadísticas $\bar{x}$ y $S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$.

Teorema 1. *Sea* $X \in \Phi_{\alpha,\sigma^2}$. *Entonces*

1) $(\bar{x} - \alpha)\sqrt{n}/\sigma \in \Phi_{0,1}$,
2) $(n-1)S_0^2/\sigma \in \mathbf{H}_{n-1}$,
3) *las variables aleatorias* $\bar{x}$ y S_0^2 *son independientes.*

Demostración. La afirmación 1 es evidente. Además está claro que sin limitar la generalidad podemos considerar $\alpha = 0$, $\sigma = 1$. Tenemos

$$(n-1)S_0^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Notemos que

$$\sqrt{n}\bar{x} = \frac{1}{\sqrt{n}} x_1 + \ldots + \frac{1}{\sqrt{n}} x_n$$

y que el vector columna n-dimensional $\begin{matrix} 1/\sqrt{n} \\ \vdots \\ 1/\sqrt{n} \end{matrix}$ (su norma vale 1) siempre puede ser completado hasta cualquier matriz ortogonal C. Entonces $y_1 = \sqrt{n}\bar{x}$ es la primera coordenada $Y = XC$ y, en virtud del lema 2, obtenemos que

$$(n-1)S_0^2 = \sum_{i=1}^{n} x_i^2 - y_1^2 \in \mathbf{H}_{n-1}$$

y que las variables aleatorias $(n-1)S_0^2$ e $y_1 = \sqrt{n}\bar{x}$ son independientes. ◁

Corolario 1. *Sea* $X \in \Phi_{\alpha,\sigma^2}$. *Entonces la variable aleatoria* $t = (\bar{x} - \alpha)\sqrt{n}/S_0 \in \mathbf{T}_{n-1}$, *o sea, tiene una distribución de Student con* $n-1$ *grados de libertad.*

Esto se deduce del teorema 1 y de la representación

$$t = \frac{(\bar{x} - \alpha)\sqrt{n}}{\sigma} \cdot \frac{1}{\sqrt{\dfrac{1}{n-1} \cdot \dfrac{(n-1)S_0^2}{\sigma^2}}}. \quad \triangleleft$$

La afirmación del teorema 1 acerca de la independencia de S_0^2 y $\bar{x}$ puede ser amplificada. Resulta que $\bar{x}$ no depende del vector $X - \bar{x}$ (o sea, que no depende de los sumandos de S_0^2). Esto se deduce de la normalidad de $\bar{x}$ y de $X - \bar{x}$, así como de la no correlatividad de las variables aleatorias $\bar{x}$ y $x_i - \bar{x}$, la cual se desprende de la igualdad ($\alpha = 0$)

$$\mathbf{M}(x_1 - \bar{x})\bar{x} = \frac{1}{n^2}\left[(n-1)\mathbf{M}x_1^2 - \mathbf{M}\left(\sum_{i=2}^{n} x_i\right)^2\right] = 0.$$

2. Construcción de intervalos confidenciales exactos para los parámetros de distribución normal. Examinemos primeramente dos situaciones elementales.

a) Supongamos que $X \in \Phi_{\alpha,\sigma^2}$ y que σ^2 se conoce. Es preciso construir el intervalo confidencial de nivel $1 - \varepsilon$ para el parámetro α. En este caso la forma del intervalo confidencial se deduce, evidentemente, de las igualdades

$$\mathbf{P}(|\overline{x} - \alpha)\sqrt{n}/\sigma| < \beta) = \mathbf{P}(-\sigma\beta/\sqrt{n} < \overline{x} - \alpha < \sigma\beta/\sqrt{n} = 1 - \varepsilon,$$

donde, como antes, $\beta = \lambda_{\varepsilon/2}$, $\Phi_{0,1}((-\infty, \lambda_\delta)) = 1 - \delta$, así que

$$\alpha^*(\varepsilon, X) = \overline{x} \pm \sigma\beta/\sqrt{n}.$$

Proponemos que el lector, en forma de ejercicio, haga uso de un procedimiento un poco más formal, expuesto en el teorema 31.2, con la utilización de la función $G(\alpha, X) = (\overline{x} - \alpha)\sqrt{n}/\sigma \in \Phi_{0,1}$.

b) Ahora supongamos que se conoce α. Es necesario construir el intervalo confidencial de nivel $1 - \varepsilon$ para σ^2.

Pongamos

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2.$$

Es evidente que en este caso $nS_1^2/\sigma^2 \in \mathbf{H}_n$ y, por consiguiente,

$$\mathbf{P}(y_n^- < nS_1^2/\sigma^2 < y_n^+) = \mathbf{H}_n((y_n^-, y_n^+)) = \mathbf{P}(nS_1^2/y_n^+ < \sigma^2 < nS_1^2/y_n^-).$$

Ahora bien, las fronteras del intervalo confidencial de nivel $1 - \varepsilon$ tendrán la forma

$$(\sigma^2)^* = nS_1^2/y_n^*$$

para todos y_n^* tales que $\mathbf{H}_n((y_n^-, y_n^+)) = 1 - \varepsilon$.

Si se utiliza el procedimiento del teorema 31.2, conviene poner $G(\sigma, X) = nS_1^2/\sigma^2 \in \mathbf{H}_n$.

Pasemos ahora al caso cuando ambos parámetros α y σ^2 se desconocen.

c) Con el fin de construir el intervalo confidencial para σ^2, hagamos uso de la estadística $G_1(\sigma, X) = (n-1)S_0^2/\sigma^2$. En virtud del teorema 1, $G_1(\sigma, X) \in \mathbf{H}_{n-1}$. Luego procedemos al igual que en el caso b). Las fronteras del intervalo confidencial para σ^2 tendrán la forma

$$(\sigma^2)^* = (n-1)S_0^2/y_{n-1}^*.$$

Es fácil ver que las estadísticas $G(\sigma, X)$ y $G_1(\sigma, X)$ en los casos b) y c) tienen la misma distribución y, por lo tanto, dan los mismos intervalos confidenciales para σ^2 siempre que en el caso b) tengamos una observación más que en el caso c). Hablando figuradamente, en el caso c) "perdemos" una observación debido a la existencia de una indeterminación adicional,

o sea, del parámetro desconocido α. Esta observación se destina, en cierto sentido, a estimar el parámetro "obstaculizante"*) α.

d) Construyamos ahora el intervalo confidencial para α. Hagamos uso de la estadística $G_1(\alpha, X) = (\overline{x} - \alpha)\sqrt{n}/S_0$. En virtud del corolario del teorema 1,

$$G_1(\alpha, X) \in \mathbf{T}_{n-1}.$$

En vista de que la función $G_1(\alpha, X)$ satisface las condiciones del teorema 31.2, los razonamientos ulteriores repiten exactamente los correspondientes razonamientos en los casos a), b) y c). Las fronteras del intervalo confidencial tienen la forma (para simplificar la exposición tomamos un intervalo simétrico)

$$\alpha^{\pm} = \overline{x} \pm \tau_\varepsilon S_0/\sqrt{n}.$$

donde τ_ε se determina de la igualdad

$$\mathbf{P}(|t_{n-1}| < \tau_\varepsilon) = \mathbf{T}_{n-1}((-\tau_\varepsilon, \tau_\varepsilon)) = 1 - \varepsilon.$$

Nótese que si el valor de S_0 es próximo al de σ, entonces el intervalo confidencial obtenido será más ancho que el dado en a), ya que $\tau_\varepsilon > \beta$ (véase la observación en el § 2). Esto se explica, como antes, por la existencia del parámetro "obstaculizante" α, el cual se conoce en a).

Los números y^*, para los cuales en las investigaciones citadas se ha cumplido la relación

$$\mathbf{P}(G(\theta, X) \in (y^-, y^+)) = 1 - \varepsilon,$$

en la práctica suelen determinarse con ayuda de las tablas de la estadística matemática.

En el § 3.8 mostraremos que los intervalos confidenciales construidos en este párrafo son, desde cierto punto de vista, los mejores.

---

*) Es interesante notar que, a pesar de las ideas intuitivas iniciales, por una observación $x_1 \in \Phi_{\alpha, \sigma^2}$ es posible construir el intervalo confidencial para σ^2, siendo α desconocido. Los siguientes razonamientos que muestran esto fueron comunicados a nosotros por L. N. Bolshakov.

Escojamos u de modo que $\Phi(1/u) - \Phi(-1/u) = \varepsilon$, donde $\Phi(x) = \Phi_{0,1}((-\infty, x))$. Entonces

$$\mathbf{P}(\sigma > u|x_1|) = \mathbf{P}(-\sigma/u < x_1 < \sigma/u) = \mathbf{P}\left(-\frac{1}{u} - \frac{\alpha}{\sigma} < \frac{(x_1 - \alpha)}{\sigma} < \frac{1}{u} - \frac{\alpha}{\sigma}\right) =$$

$$= \Phi\left(\frac{1}{u} - \frac{\alpha}{\sigma}\right) - \Phi\left(-\frac{1}{u} - \frac{\alpha}{\sigma}\right) \leqslant \Phi\left(\frac{1}{u}\right) - \Phi\left(-\frac{1}{u}\right) = \varepsilon. \triangleleft$$