Rev.R.Acad.Cienc.Exact.Fis.Nat. (Esp),
Vol. 95, N.os 1–2, pp 81–99, 2002
Monográfico: Tricentenario de Thomas Bayes

UN PROGRAMA DE SÍNTESIS PARA LA ENSEÑANZA UNIVERSITARIA DE LA ESTADÍSTICA MATEMÁTICA CONTEMPORÁNEA

(Concepto de probabilidad/Contraste de hipótesis/Distribuciones de referencia/Distribuciones en el muestreo/Estimación/Función de pérdida/Función de verosimilitud/Intercambiabilidad/Métodos Bayesianos objetivos/Modelos probabilísticos/Simulación/Teoría de la decisión)

JOSÉ MIGUEL BERNARDO HERRÁNZ *

* Departament d'Estadística i I. O., Universitat de València. Facultat de Matemàtiques, 46100–Burjassot. España

ABSTRACT

In this paper it is argued that the current conventional approach to the education in theoretical statistics of university students requires a radical reformulation. A possible alternative syllabus is presented which could be used with students that, with an appropriate mathematical basis (basic calculus and probability theory) are subject to a primer in mathematical statistics. More specifically, the proposed syllabus contains an elementary introduction to decision theory, a description of probability models based on the concept of exchangeability and the study of the likelihood function, an introduction to objective Bayesian methods and an analysis of the behaviour under repeated sampling of statistical methods which includes, and reformulates, the more relevant concepts of frequentist statistics.

RESUMEN

En este trabajo se defiende la necesidad de una profunda reconversión en la enseñanza universitaria de la estadística matemática y se apunta el contenido de un programa que podría resultar apropiado para los alumnos que, con una base matemática apropiada (fundamentos de cálculo y elementos de teoría de la probabilidad), se enfrenten por primera vez a los conceptos básicos de esta disciplina. En particular, el programa propuesto incluye una introducción elemental a la teoría de la decisión, una descripción de los modelos probabilísticos basada en el concepto de intercambiabilidad y en el estudio de la función de verosimilitud, una introducción a los métodos bayesianos objetivos, y un análisis del comportamiento bajo muestreo repetido de los métodos estadísticos, que recoge y reinterpreta los conceptos más relevantes de los métodos frecuentistas.

1 INTRODUCCIÓN

Un análisis comparado de los programas en los que se basa la enseñanza universitaria de la estadística matemática en los primeros ciclos de la enseñanza universitaria de casi todo el mundo pone de manifiesto un notable desfase entre lo que se enseña como 'teoría establecida' y lo que se investiga y (sobre todo) se aplica en el mundo real. En efecto, los programas en vigor son una réplica (casi sin modificaciones) de los programas que, basados en un paradigma frecuentista, se enseñaban ya a mediados del siglo pasado. Lamentablemente, tales programas (i) frecuentemente ignoran los numerosos problemas, contraejemplos y limitaciones asociados al uso exclusivo de ese paradigma, y (ii) típicamente eluden cualquier comentario sobre el paradigma bayesiano. Se sabe, sin embargo, que un importante porcentaje de la investigación publicada en las revistas profesionales de primera línea se sitúa dentro del paradigma bayesiano, y es ya de dominio público que las aplicaciones más espectaculares de la estadística matemática en los últimos años (desde la bioinformática a la teledetección, desde la traducción automática al diagnóstico médico automatizado, desde la gestión de una cartera de valores al análisis de datos en la arqueología moderna) se realizan preferentemente desde un paradigma bayesiano (y frecuentemente no es posible hacerlo de otra manera). El aprendizaje de los métodos estadísticos más modernos queda así relegado a asignaturas optativas en el segundo o en el tercer ciclo de algunos estudios, con lo que la inmensa mayoría de los estudiantes universitarios (que

únicamente tienen una asignatura de estadística en el primer ciclo) reciben una formación en estadística matemática definitivamente incompleta, resultando asi privados de conocimientos muy relevantes para su práctica profesional.

Históricamente, siempre ha existido un cierto retraso en la incorporación a la enseñanza universitaria de los nuevos paradigmas científicos (el caso de la física relativista es un ejemplo no muy lejano); sin embargo, este factor de inercia no es por si mismo suficiente para explicar el retraso observable en la incorporación del paradigma bayesiano a la enseñanza universitaria de la estadística matemática. En las numerosas ocasiones en las que este debate se ha planteado en distintos foros suelen escucharse dos tipos de argumentos por parte de quienes prefieren mantener la situación actual: (i) los métodos bayesianos son necesariamente subjetivos, y resultan por lo tanto inadecuados en el análisis de la investigación científica, y (ii) los alumnos tienen que conocer los métodos frecuentistas, por que van a necesitarlos en su vida profesional, y no es posible integrar ambos paradigmas de forma comprensible en una única asignatura.

El primer argumento sólo pone de manifiesto la falta de información de quienes lo sostienen. Las soluciones bayesianas *objetivas* a los problemas estadísticos que pueden abordarse en un primer curso de estadística matemática (objetivas en el sentido de que, como las soluciones frecuentistas, sólo dependen del modelo asumido y de los datos observados) se conocen desde finales de los años 60 (Lindley, 1965; Zellner, 1971; Press, 1972; Box & Tiao, 1973). Además, el análisis de referencia, desarrollado en los 80's y los 90's (Bernardo, 1979; Berger & Bernardo, 1992; Bernardo & Smith, 1994) hace tiempo que ha proporcionado una solución general que incluye, sistematiza y generaliza, tales soluciones.

El segundo argumento tiene mucho mayor calado. Es innegable que cualquier profesional que utilice métodos estadísticos debería conocer los métodos frecuentistas, no sólo porque su uso es todavía dominante, sino también porque pueden proporcionar valiosa información sobre el comportamiento que puede esperarse de *cualquier* metodología. Por otra parte, son asimismo innegables las dificultades inherentes al intentar describir dentro de una misma asignatura dos paradigmas que suelen ser presentados como incompatibles. El objeto de este trabajo es proponer un *programa de síntesis* en el que los métodos frecuentistas, cuya *raison d'être* es el análisis del comportamiento esperado bajo muestreo repetido de *cualquier* procedimiento estadístico, son precisamente utilizados para analizar el comportamiento esperado bajo muestreo repetido de los métodos bayesianos objetivos, demostrando que gozan de propiedades frecuentistas especialmente atractivas, que permiten darles una interpretación dual. Para citar el ejemplo más conocido, dada una muestra normal de tamaño n, con media $\bar{x}$ y desviación típica s, el intervalo $\bar{x} \pm t_{\alpha/2}\, s/\sqrt{n-1}$ se obtiene desde la perspectiva bayesiana objetiva como una región de confianza a la que el verdadero valor de la media μ de una población normal pertenece con una *credibilidad racional* $1-\alpha$ (en una escala $[0,1]$), que es precisamente el tipo de resultado (condicional a los datos realmente observados) que obviamente interesa en cualquier aplicación concreta. Por otra parte, el análisis del comportamiento en el muestreo de esa región demuestra que, bajo uso repetido, tal intervalo *contendría* al verdadero valor de μ en el $100(1-\alpha)\%$ de los casos, proporcionando así una valiosa calibración del resultado bayesiano. La correspondencia entre los conceptos de región de confianza creíble y de región de confianza frecuentista (exacta en este ejemplo), resulta ser casi siempre una aproximación válida para muestras suficientemente grandes.

El programa que proponemos tiene cinco partes: *Fundamentos*, *Modelos Probabilísticos*, *Inferencia*, *Calibración* y *Aplicaciones*, cuyo contenido es descrito en las próximas secciones. Para mantener este trabajo dentro de unas dimensiones razonables, tan sólo analizaremos con algún detalle los elementos menos frecuentes en una presentación frecuentista de los métodos estadísticos, evitando en lo posible el enunciado de definiciones o resultados típicamente incluidos en un libro de texto convencional. La sección final contiene una descripción abreviada de la propuesta presentada, articulada en un posible *programa* para una asignatura de *Estadística Matemática* que se integre en un *primer ciclo* universitario, y en el que se consideran tanto los aspectos teóricos como su implementación práctica con un software adecuado.

El desarrollo de un programa de síntesis como el propuesto plantea importantes problemas de notación; se ha optado por una notación en la que se prima la sencillez. En particular, se asume que todas las distribuciones de probabilidad pueden ser descritas mediante sus funciones de probabilidad o de densidad de probabilidad y, generalmente, se utiliza la misma notación para una variable aleatoria y para los valores que puede adoptar. Se utiliza el alfabeto latino para variables aleatorias *observables* (típicamente datos) y para constantes conocidas, y el alfabeto griego para variables aleatorias *inobservables* (típicamente parámetros), utilizándose negritas cuando se trata de vectores. Se utilizan letras minúsculas para las variables y mayúsculas para sus dominios. Se hace uso de la convención matemática que permite referirse a las funciones f y g de $\boldsymbol{x} \in \mathcal{X}$ mediante $f(\boldsymbol{x})$ y $g(\boldsymbol{x})$ respectivamente. En particular, $p(\boldsymbol{x} \mid C)$, $p(\boldsymbol{y} \mid C)$, generalmente representan densidades de probabilidad de los vectores observables $\boldsymbol{x}$ e $\boldsymbol{y}$ en condiciones C, sin que esto sugiera que $\boldsymbol{x}$ e $\boldsymbol{y}$ tienen la misma distribución; análogamente, $\pi(\boldsymbol{\theta} \mid D)$ y $\pi(\boldsymbol{\omega} \mid D)$ generalmente representan densidades de probabilidad de los vectores *inobservables* $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$, dada la información proporcionada por D. Si alguno de estos vectores aleatorios es dis-

creto, se utiliza la misma notación para representar su correspondiente función de probabilidad. En aquellas ocasiones en las que resulte conveniente distinguir entre la función de densidad y su valor en un punto, utilizaremos respectivamente notaciones del tipo $p_{\boldsymbol{x}}(.)$ y $p_{\boldsymbol{x}}(\boldsymbol{x})$ o $\pi_{\boldsymbol{\theta}}(.)$ y $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. De forma general, representaremos mediante $\boldsymbol{z} \in \mathcal{Z}$ al conjunto de todos los *datos disponibles*, cualquiera que sea su estructura; por ejemplo, $\Pr(\boldsymbol{\theta} \in A \,|\, \boldsymbol{z}, C) = \int_A \pi(\boldsymbol{\theta} \,|\, \boldsymbol{z}, C)\, d\boldsymbol{\theta}$ representa la probabilidad de que $\boldsymbol{\theta}$ pertenezca a la región $A \subset \Theta$, dados los datos $\boldsymbol{z}$ y las condiciones C. Finalmente, funciones específicas de probabilidad o de densidad de probabilidad son representadas con nombres apropiados. Por ejemplo, si x es normal con media μ y varianza σ^2, su densidad de probabilidad se escribe $\mathrm{N}(x \mid \mu, \sigma^2)$; si θ tiene una distribución Beta con parámetros α y β, su densidad de probabilidad se escribe $\mathrm{Be}(\theta \,|\, \alpha, \beta)$.

2 FUNDAMENTOS

En la primera parte del programa propuesto se desarrollan, a nivel muy elemental, los fundamentos sobre los que descansa la estadística matemática moderna. En particular, (i) se revisa el *concepto de probabilidad*, insistiendo en su doble interpretación (como medida racional del grado de creencia en la ocurrencia de un suceso incierto condicionada a la información disponible, y como límite de frecuencias relativas), y se repasan los conceptos básicos de la teoría de la probabilidad; (ii) se introducen nociones elementales de la *teoría de la decisión*, lo que proporciona una estructura teórica que permite ordenar y sistematizar muchos de los conceptos posteriores; finalmente, (iii) se introducen funciones de pérdida especialmente útiles para la formulación de los problemas básicos de la estadística matemática, con especial énfasis en las *medidas de discrepancia entre distribuciones* basadas en la teoría matemática de la información.

2.1 Medida de Probabilidad

En este primer tema se pone de manifiesto que la interpretación del concepto de probabilidad que va a utilizarse responde exactamente al uso epistemológico del concepto de probabilidad en el lenguaje cotidiano. Se insiste en el hecho de que tal interpretación resulta necesaria para poder abordar las numerosas aplicaciones reales en las que no existen 'repeticiones bajo condiciones idénticas' (predicciones electorales, diagnóstico médico automatizado, marketing, . . .).

Se describe, sin demostraciones, como una formulación precisa de las propiedades exigibles a una medida racional de la credibilidad de los sucesos inciertos permite demostrar que tal medida debe tener precisamente la estructura matemática de una medida de probabilidad.

Texto de la Pregunta

$\boxed{p_1}$	R_1=Respuesta 1
$\boxed{p_2}$	R_2=Respuesta 2
	. . . . .
$\boxed{p_k}$	R_k=Respuesta k

Figura 1. Estructura de una pregunta con un conjunto de k respuestas posibles, mutuamente excluyentes, una de las cuales es correcta. La contestación del alumno debe ser una distribución de probabilidad (grado racional de creencia) $\{p_1, \ldots, p_k\}$ sobre el conjunto $\{R_1, \ldots, R_k\}$ de las respuestas propuestas..

Se utilizan los cuestionarios de respuesta múltiple para ilustrar (en un entorno bien conocido por los alumnos) este concepto de probabilidad, destacándose que la única respuesta razonable a una pregunta con k respuestas posibles mutuamente excluyentes (con una estructura como la de la Figura 1) es una *distribución de probabilidad* (que puede ser degenerada) $\boldsymbol{p} = \{p_1, \ldots, p_k\}$ que describa el grado de creencia del alumno en cada una de las respuestas. Por ejemplo, un alumno que dude entre las dos primeras respuestas, pero que descarte todas las demás, podría contestar especificando la distribución $\boldsymbol{p} = \{0.5, 0.5, 0, \ldots, 0\}$.

Para motivar el concepto de la utilidad asociada a una predicción probabilística $\boldsymbol{p}$ (que se discute en el tema siguiente), se plantea el problema de la calificación $U(\boldsymbol{p}, R_t)$ que debe merecer una respuesta probabilística

$$\boldsymbol{p} = \{p_1, p_2, \ldots, p_k\}, \quad p_i \geq 0, \quad \sum\nolimits_{j=1}^{k} p_j = 1,$$

en función del resultado correcto, R_t.

Proporcionando ejemplos relevantes, se demuestra que la interpretación de la probabilidad como grado de creencia racional incluye, como casos particulares, la interpretación clásica, basada en simetrías postuladas, y la interpretación frecuentista, basada en el límite de frecuencias relativas de sucesos que (supuestamente) tienen lugar en condiciones idénticas.

Se repasan (y se ilustran con ejemplos) los conceptos básicos de teoría de probabilidad que van a resultar necesarios en la asignatura: cantidades y vectores aleatorios, funciones de distribución, funciones de probabilidad y de densidad de probabilidad, distribuciones marginales y condicionales, esperanza matemática, momentos, cuantiles, y teoremas de cambio de variable.

Presentándolo como una aplicación de los teoremas de cambio de variable, el tema concluye con la descripción del algoritmo básico de *simulación* de observaciones procedentes de una distribución conocida por transformación (mediante la función inversa de su función de distribución) de observaciones uniformes en el intervalo unidad.

2.2 Introducción a la Teoría de Decisión

La teoría de la decisión proporciona la estructura adecuada para comparar entre si distintos procedimientos estadísticos.

En este tema se describe la estructura básica de un problema de decisión en ambiente de incertidumbre: el espacio de *acciones alternativas* $\mathcal{A} = \{a_i, i \in I\}$, los conjuntos de *sucesos inciertos relevantes* $\{\Theta_i, i \in I\}$, con $\Theta_i = \{\theta_{ij}, j \in J_i\}$, y el conjunto de *consecuencias* posibles $\mathcal{C} = \{c_{ij} = (a_i, \theta_{ij}), i \in I,\ j \in J_i\}$.

Se establece una analogía con la geometría euclídea para establecer un conjunto de *postulados* que definan el concepto de decisión racional, y se argumentan las ventajas de disponer de un *sistema axiomático* que, como en la geometría euclídea, permita la deducción coherente de resultados.

Se describe, sin intentarse una definición formal, el contenido intuitivo de los axiomas básicos de *comportamiento coherente* (comparabilidad, transitividad, consistencia y existencia de sucesos estándard) y se enuncian sus consecuencias fundamentales: para tomar decisiones de forma coherente, es necesario:

(i) Establecer *medidas de crediblidad* para los distintos sucesos inciertos relevantes, que deben tener la estructura matemática de distribuciones de probabilidad condicionales a los datos disponibles $\boldsymbol{z}$, de forma que, $\forall i \in I$,

$$\exists\ \Pr(\theta_{ij} \mid a_i, \boldsymbol{z}) \geq 0;\ \sum\nolimits_{j \in J_i} \Pr(\theta_{ij} \mid a_i, \boldsymbol{z}) = 1.$$

(ii) Establecer una *función de utilidad* $U(a_i, \theta_{ij}) \in \Re$, de forma que $U(a_i, \theta_{ij})$ que describa, en unidades apropiadas, la deseabilidad de la consecuencia $c_{ij} = (a_i, \theta_{ij})$ a la que conduciría la alternativa a_i si tuviese lugar el suceso θ_{ij}.

(iii) Medir el valor esperado de una alternativa a_i mediante su utilidad esperada $\overline{U}(a_i \mid \boldsymbol{z})$, definida como la *media* de las utilidades $U(a_i, \theta_{ij})$ a que podría dar lugar, *ponderada* por sus respectivas probabilidades $\Pr(\theta_{ij} \mid a_i, \boldsymbol{z})$, de forma que

$$\overline{U}(a_i \mid \boldsymbol{z}) = \sum\nolimits_{j \in J_i} U(a_i, \theta_{ij}) \Pr(\theta_{ij} \mid a_i, \boldsymbol{z}).$$

Finalmente, se introduce el concepto de *función de pérdida* $L(a_i, \theta_{ij})$, definida como la pérdida de utilidad que se sufriría eligiendo a_i si sucediese θ_{ij}, en lugar de la acción $a_j^* = \sup_{k \in I}\{U(a_k, \theta_{ij})\}$ que hubiera resultado óptima en ese caso, es decir

$$L(a_i, \theta_{ij}) = U(a_j^*, \theta_{ij}) - U(a_i, \theta_{ij}).$$

Se demuestra, que con esta definición, la acción que maximiza la utilidad esperada es necesariamente la misma acción que minimiza la pérdida esperada.

El tema concluye reformulando los resultados anteriores para el caso en el que los sucesos inciertos relevantes son vectores aleatorios continuos de forma que, por ejemplo, la acción óptima $a^*(\boldsymbol{z})$ viene dada por

$$\begin{aligned} a^*(\boldsymbol{z}) &= \arg\max_{a_i \in \mathcal{A}} \int_\Theta U(a_i, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta} \\ &= \arg\min_{a_i \in \mathcal{A}} \int_\Theta L(a_i, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta}. \end{aligned}$$

2.3 Medidas de Información y de Discrepancia

Se retoma el ejemplo proporcionado por los cuestionarios de respuesta múltiple para introducir el concepto de la utilidad $U(p_x(.), x)$ de la predicción probabilística proporcionada por una función de probabilidad $p_x(.)$ sobre los posibles valores de una cantidad aleatoria observable en función del valor x finalmente observado. Se demuestra que, bajo hipótesis razonables, una solución apropiada resulta ser de la forma

$$U(p_x(.), x_i) = A \log p_x(x_i) + B, \quad A > 0.$$

En particular, se describe el comportamiento de la función

$$U(\{p_1, \ldots, p_k\}, R_t) = 1 - \frac{\log(p_t)}{\log(1/k)}, \quad p_i = \Pr(R_i),$$

para especificar la calificación merecida por una respuesta $\{p_1, \ldots, p_k\}$ en función del resultado correcto R_t. Se comprueba que esta función asocia el valor unidad a una respuesta perfecta ($p_t = 1, p_j = 0, \forall j \neq t$) y el valor cero a la falta total de información, convencionalmente descrita por una distribución uniforme $\boldsymbol{p} = \{1/k, \ldots, 1/k\}$, con lo que (necesariamente) asocia valores negativos a los disparates (las respuestas que asignan probabilidades altas a respuestas incorrectas).

Formalmente, el incremento de utilidad que puede *esperarse* de una predicción $\{p_1, \ldots, p_k\}$, sobre el valor de una variable aleatoria x con un número finito $\{x_1, \ldots, x_k\}$ de valores posibles, $p_i = \Pr(x = x_i)$, definido como el incremento de utilidad que puede esperarse de una predicción $\{p_1, \ldots, p_k\}$ por encima de una predicción trivial $\{1/k, \ldots, 1/k\}$, resulta ser:

$$\begin{aligned} &\sum_{i=1}^{k} [U(\{p_1, \ldots, p_k\}, x_i) - U(\{1/k, \ldots, 1/k\}, x_i)]\ p_i \\ &\qquad = A \sum_{i=1}^{k} p_i \log \frac{p_i}{1/k}, \quad A > 0, \end{aligned}$$

lo que constituye una medida de la *cantidad de información* sobre el valor de la variable aleatoria discreta x que contiene la predicción $\{p_1, \ldots, p_k\}$, tomando como origen la distribución uniforme $\{1/k, \ldots, 1/k\}$.

En general, dado un vector aleatorio $\boldsymbol{x} \in \mathcal{X}$, la cantidad de información sobre $\boldsymbol{x}$ contenida en una densidad de probabilidad $p_{\boldsymbol{x}}(\cdot)$, tomando como origen otra densidad $q_{\boldsymbol{x}}(\cdot)$, se define como

$$K\{q_{\boldsymbol{x}}(.) \,|\, p_{\boldsymbol{x}}(.)\} = \int_{\mathcal{X}} p_{\boldsymbol{x}}(\boldsymbol{x}) \log \frac{p_{\boldsymbol{x}}(\boldsymbol{x})}{q_{\boldsymbol{x}}(\boldsymbol{x})}\, d\boldsymbol{x},$$

una cantidad no-negativa e invariante ante transformaciones biyectivas de $\boldsymbol{x}$, conocida como la *divergencia logarítmica* de $q_{\boldsymbol{x}}(\cdot)$ a $p_{\boldsymbol{x}}(\cdot)$. Se describe cómo la teoría matemática de la información proporciona una interpretación operativa de la divergencia logarítmica introducida en términos de las unidades de información (*bits* si se utilizan logaritmos de base 2) necesarias para obtener $p_{\boldsymbol{x}}(\cdot)$ a partir de $q_{\boldsymbol{x}}(\cdot)$. Se subraya que *no* se trata de una función simétrica de forma que, en general, $K\{q \,|\, p\} \neq K\{p \,|\, q\}$.

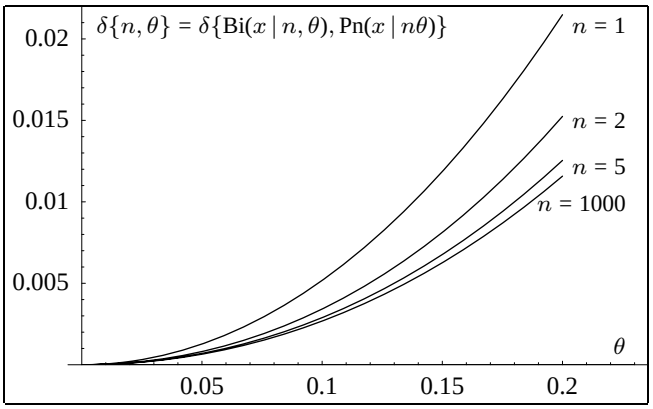

Figura 2. Discrepancia intrínseca $\delta(n,\theta)$ entre una distribución Binomial $\text{Bi}(x \mid n,\theta)$ y una Poisson $\text{Pn}(x \mid n\theta)$ como función de θ, para $n = 1, 2, 5$ y 1000.

La divergencia logarítmica permite definir una medida general de discrepancia (simétrica) entre dos distribuciones de probabilidad, la *discrepancia intrínseca*, definida por

$$\delta\{p_{\boldsymbol{x}}(.), q_{\boldsymbol{x}}(.)\} = \min\left[K\{p_{\boldsymbol{x}}(.) \,|\, q_{\boldsymbol{x}}(.)\},\ K\{q_{\boldsymbol{x}}(.) \,|\, p_{\boldsymbol{x}}(.)\}\right].$$

Por ejemplo, la Figura 2 representa la discrepancia intrínseca entre una distribución binomial $\text{Bi}(x \mid \theta, n)$ y una Poisson $\text{Pn}(x \mid n\theta)$, poniéndose de manifiesto que la condición realmente importante para que la aproximación funcione bien es que θ sea pequeño.

Si $\boldsymbol{x} \in \mathcal{X}$ es un vector aleatorio de cuya función de densidad de probabilidad se sabe que es $p_1(\boldsymbol{x})$ o $p_2(\boldsymbol{x})$, entonces la discrepancia intrínseca $\delta\{p_1, p_2\}$ es el mínimo valor esperado del logaritmo del cociente de densidades

$$\log\left[\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\right]$$

en favor de la densidad verdadera. En particular, si $p_1(\boldsymbol{z})$ y $p_2(\boldsymbol{z})$ son modelos alternativos para un conjunto de datos $\boldsymbol{z} \in \mathcal{Z}$ y se supone que uno de ellos es cierto, entonces $\delta\{p_1, p_2\}$ es el mínimo valor esperado del logaritmo del cociente de verosimilitudes en favor del modelo verdadero; este resultado es importante en el contexto del problema de contraste de hipótesis (Sección 4.4).

La discrepancia intrínseca permite definir un tipo de convergencia entre distribuciones de probabilidad especialmente útil en estadística matemática, la *convergencia intrínseca*. Se dice que una sucesión de densidades de probabilidad (funciones de probabilidad) $\{p_i\}_{i=1}^{\infty}$ converge intrínsecamente a una densidad (función de probabilidad) p cuando la sucesión (de números reales positivos) $\delta\{p_i, p\}_{i=1}^{\infty}$ converge a cero:

$$\delta \lim p_i = p \iff \lim_{i\to\infty} \delta\{p_i, p\} = 0.$$

La *dependencia intrínseca* $D(p_{\boldsymbol{xy}})$ entre dos vectores aleatorios, $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$ con densidad de probabilidad conjunta $p_{\boldsymbol{xy}}(\cdot,\cdot)$ se define como la discrepancia intrínseca entre su distribución conjunta y el producto $p_{\boldsymbol{x}}(\cdot)\, p_{\boldsymbol{y}}(\cdot)$ de sus distribuciones marginales, esto es

$$\begin{aligned} D(p_{\boldsymbol{xy}}) &= \delta\{p_{\boldsymbol{xy}}, p_{\boldsymbol{x}}p_{\boldsymbol{y}}\} \\ &= \min\{K\{p_{\boldsymbol{x}}p_{\boldsymbol{y}} \,|\, p_{\boldsymbol{xy}}\}, K\{p_{\boldsymbol{xy}} \,|\, p_{\boldsymbol{x}}p_{\boldsymbol{y}}\}\} \\ &= K\{p_{\boldsymbol{x}}p_{\boldsymbol{y}} \,|\, p_{\boldsymbol{xy}}\} \\ &= \int_{\mathcal{X}\mathcal{Y}} p_{\boldsymbol{xy}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{p_{\boldsymbol{xy}}(\boldsymbol{x},\boldsymbol{y})}{p_{\boldsymbol{x}}(\boldsymbol{x})p_{\boldsymbol{y}}(\boldsymbol{y})}\, d\boldsymbol{x}d\boldsymbol{y}, \end{aligned}$$

(con sumas en lugar de integrales en el caso discreto), y proporciona una medida *general* (no-negativa e invariante frente a biyecciones) de la dependencia entre dos vectores aleatorios que se anula si, y sólamente si, los vectores son independientes. Cuando la distribución conjunta es normal bivariante se obtiene una sencilla función del coeficiente de correlación lineal:

$$\begin{aligned} D\left\{\text{N}_2\left\{ \begin{pmatrix} x \\ y \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right\}\right\} \\ = \log\left[\frac{1}{\sqrt{1-\rho^2}}\right]. \end{aligned}$$

Finalmente, la discrepancia entre dos conjuntos de distribuciones de probabilidad se define como la discrepancia intrínseca *mínima* entre elementos de ambos conjuntos. En particular, la discrepancia intrínseca

$\delta(M_1, M_2)$ entre dos familias de modelos probabilísticos paramétricos definidos para $\boldsymbol{x} \in \mathcal{X}$,

$$M_1 = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}, \quad M_2 = \{q(\boldsymbol{x} \mid \boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi\},$$

viene dada por

$$\delta\{M_1, M_2\} = \inf_{\boldsymbol{\theta}\in\Theta,\, \boldsymbol{\phi}\in\Phi} \delta\{p_{\boldsymbol{x}}(. \mid \boldsymbol{\theta}), q_{\boldsymbol{x}}(. \mid \boldsymbol{\phi})\},$$

lo que, como se verá mas adelante, encuentra aplicaciones inmediatas en la formalización de los problemas convencionales de *estimación puntual* y de *contraste de hipótesis*.

3 MODELOS PROBABILÍSTICOS

En la segunda parte del programa propuesto se introduce el concepto básico de *intercambiabilidad,* lo que permite una definición formal de modelo probabilístico y una definición operativa de sus parámetros. Se introduce el concepto de *muestra aleatoria*, se utiliza el procedimiento de simulación descrito en la primera parte para simular muestras aleatorias de un modelo probabilístico de parámetros conocidos, y se introduce el concepto de *distribución en el muestreo*. Se define y se estudia con detalle la *función de verosimilitud*. Finalmente, se define el concepto de *suficiencia*, y se introduce la *familia exponencial de distribuciones*.

3.1 Intercambiabilidad e Independencia Condicional

En este tema se formaliza la idea intuitiva de observaciones 'homogéneas' que resulta central a cualquier intento de razonamiento inferencial: un conjunto de vectores aleatorios $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ es *intercambiable* si su distribución conjunta es invariante ante permutaciones, de forma que el orden en el que son observados los vectores no proporciona información alguna sobre su comportamiento. Una sucesión de vectores aleatorios es intercambiable si cualquiera de sus subconjuntos finitos es intercambiable.

Se insiste en que los resultados experimentales suelen estar constituidos por grupos de observaciones intercambiables, pero muy rara vez independientes. En efecto, frecuentemente $p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_2, \boldsymbol{x}_1)$ (i.e. $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ son intercambiables), pero $\boldsymbol{x}_1$ típicamente proporciona información sobre $\boldsymbol{x}_2$, de forma que $p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) \neq p(\boldsymbol{x}_2)$, con lo que $\boldsymbol{x}_1$ y $\boldsymbol{x}_2$ *no* son independientes.

Se enuncia (sin demostración) el *teorema general de representación* (que, lamentablemente, no suele ser incluido en los programas de teoría de probabilidad):

Si $\{\boldsymbol{x}_j\}_{j=1}^{\infty}$ es una sucesión intercambiable de vectores aleatorios, $\boldsymbol{x}_j \in \mathcal{X}$, entonces

(i) $\exists\, g_n(\cdot);\ \boldsymbol{\theta} = \lim_{n\to\infty} g_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{\theta} \in \Theta$, (parámetro)

(ii) $\exists\, p(\boldsymbol{x} \mid \boldsymbol{\theta}) > 0, \quad \int_{\mathcal{X}} p(\boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{x} = 1$, (modelo probabilístico)

(iii) $\exists\, \pi(\boldsymbol{\theta}) > 0, \quad \int_{\Theta} \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$, (distribución inicial)

y la distribución conjunta $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de cualquier subconjunto finito $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ de la sucesión tiene una *representación integral* de la forma

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \int_{\Theta} \prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Consecuentemente, si los datos son intercambiables, entonces son *condicionalmente independientes*, de forma que

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\theta}).$$

Más precisamente, si $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ son observaciones intercambiables, entonces constituyen una *muestra aleatoria* de algún *modelo probabilístico* $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, cuyo *parámetro* $\boldsymbol{\theta}$ es el límite (cuando $n \to \infty$) de alguna función $g_n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de las observaciones. Por ejemplo, en el caso de observaciones Bernoulli, el parámetro θ se *define* como el límite de la frecuencia relativa de éxitos. De forma análoga, en el caso normal el parámetro de localización μ se *define* como el límite de la media aritmética de las observaciones.

Además, la información de que se dispone sobre el valor del parámetro $\boldsymbol{\theta}$ debe ser descrita mediante una medida de probabilidad $\pi(\boldsymbol{\theta})$, la *distribución inicial*, cuya existencia garantiza el teorema. Consecuentemente, siempre que se trabaje con observaciones intercambiables (y cualquier muestra aleatoria es necesariamente intercambiable), la frase (desgraciadamente demasiado frecuente) 'no existe una distribución inicial sobre los parámetros' es sencillamente incompatible con la teoría matemática de la probabilidad.

Se subraya que el teorema de representación es básicamente un *teorema de existencia*: demuestra la existencia de $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ y de $\pi(\boldsymbol{\theta})$, pero tanto el modelo probabilístico $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ como la distribución inicial $\pi(\boldsymbol{\theta})$ deben ser explícitamente determinados.

Se menciona la posibilidad de *caracterizar* el modelo probabilístico a partir de la intercambiabilidad y de condiciones adicionales; en particular, se cita que (suponiendo que los $\boldsymbol{x}_i$ son intercambiables):

(i) si $\boldsymbol{x}_i \in \{0,1\}^k$, entonces el modelo es multinomial (binomial para $k = 1$), y que

(ii) si $\boldsymbol{x} \in \Re^k$ y existe simetría esférica centrada, entonces el modelo es normal k-variante.

Se adelanta que la especificación de una distribución inicial que no introduzca información adicional es posible a partir a la teoría matemática de la información, mediante un algoritmo que será descrito más adelante.

Para los detalles técnicos relativos a este tema el alumno es referido tanto a las fuentes originales (de Finetti, 1937; Hewitt & Savage, 1955) como a descripciones más recientes (Lindley & Phillips, 1976; Bernardo & Smith, 1994, Cap. 4).

3.2 Función de Verosimilitud

De forma general, se define la *función de verosimilitud* $l(\boldsymbol{\theta} \mid \boldsymbol{z})$ correspondiente a un conjunto de datos $\boldsymbol{z} \in \mathcal{Z}$ cuya distribución conjunta depende de $\boldsymbol{\theta}$, como la densidad de probabilidad (función de probabilidad con datos discretos) $l(\boldsymbol{\theta} \mid \boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{\theta})$ asociada a los datos observados $\boldsymbol{z}$ como función del valor (desconocido) del vector paramétrico $\boldsymbol{\theta}$. En particular, para datos intercambiables $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\mathcal{Z} = \mathcal{X}^n$, resulta $l(\boldsymbol{\theta} \mid \boldsymbol{z}) = \prod_i p(\boldsymbol{x}_i \mid \boldsymbol{\theta})$.

Se introduce de forma intuitiva el concepto de estimador puntual como el de una función de los datos $\boldsymbol{\theta}^e = \boldsymbol{\theta}^e(\boldsymbol{z})$ que presumiblemente aproxima el verdadero (y desconocido) valor del parámetro (la definición formal de estimador se presenta en la tercera parte del programa). Se define un estimador máximo-verosímil,

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{z}) = \arg\max_{\boldsymbol{\theta} \in \Theta} l(\boldsymbol{\theta} \mid \boldsymbol{z}),$$

y se analizan, con ejemplos, las condiciones de su existencia y unicidad. Se destaca que se trata de un concepto *invariante*: el estimador máximo-verosímil $\hat{\phi}$ de cualquier transformación biyectiva del parámetro $\phi = \phi(\boldsymbol{\theta})$, es simplemente $\phi(\hat{\boldsymbol{\theta}})$.

La evidente variabilidad del estimador $\hat{\boldsymbol{\theta}}(\boldsymbol{z})$ en función de los datos observados $\boldsymbol{z} \in \mathcal{Z}$ se utiliza para motivar y definir la *distribución en el muestreo* $p(\boldsymbol{t} \mid \boldsymbol{\theta})$ de una función de los datos $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ como la distribución de $\boldsymbol{t}$ (condicional a $\boldsymbol{\theta}$) que resulta inducida por $p(\boldsymbol{z} \mid \boldsymbol{\theta})$.

Particularizando al caso univariante, se introduce la *función de soporte* ('score'), como la derivada del logaritmo de la función de verosimilitud con respecto al parámetro, considerada como función de los datos,

$$s_\theta(\boldsymbol{z}) = \frac{\partial}{\partial \theta} \log p(\boldsymbol{z} \mid \theta), \quad \boldsymbol{z} \in \mathcal{Z},$$

y se demuestra que su distribución en el muestreo tiene media cero y que, bajo condiciones de regularidad, su varianza es

$$\int_{\mathcal{Z}} p(\boldsymbol{z} \mid \theta) \left[\frac{\partial}{\partial \theta} \log p(\boldsymbol{z} \mid \theta) \right]^2 d\boldsymbol{z}$$
$$= -\int_{\mathcal{Z}} p(\boldsymbol{z} \mid \theta) \frac{\partial^2}{\partial \theta^2} \log p(\boldsymbol{z} \mid \theta)\, d\boldsymbol{z} = h(\theta),$$

donde $h(\theta)$ es la *función de información de Fisher* correspondiente a $p(\boldsymbol{z} \mid \theta)$. Consecuentemente, la función de soporte normalizada, $s_\theta(\boldsymbol{z})/\sqrt{h(\theta)}$, tiene una distribución en el muestreo de media cero y varianza unidad lo que, como se verá más adelante, resulta especialmente útil para obtener resultados inferenciales aproximados.

El concepto de distribución bajo muestreo repetido es ilustrado con ejemplos construidos con datos simulados, analizándose las distribuciones en el muestreo del estimador máximo-verosímil y de la función de soporte normalizada y las distribuciones en el muestreo de los extremos (máximo y mínimo). Se estudia asimismo la variabilidad en el muestreo de la propia función de verosimilitud.

El tema concluye justificando la observación empírica (por simulación) del frecuente comportamiento aproximadamente normal de la distribución en el muestreo del estimador máximo-verosímil.

Utilizando el desarrollo en serie de Taylor de la función de verosimilitud, $l(\theta \mid \boldsymbol{z})$, $\theta \in \Theta \subset \Re$, alrededor de $\hat{\theta}(\boldsymbol{z})$, se pone de manifiesto que, en el caso de una muestra aleatoria $\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ y bajo condiciones de regularidad apropiadas,

(i) La distribución en el muestreo de $\hat{\theta}$, el estimador máximo verosímil, converge asintóticamente a una distribución normal con media θ y precisión $n\, i(\theta)$, donde $i(\theta)$ es la función de información de Fisher correspondiente a $p(\boldsymbol{x} \mid \theta)$.

(ii) La distribución en el muestreo de $s_\theta(\boldsymbol{z})/\sqrt{n\, i(\theta)}$, la función de soporte normalizada, converge asintóticamente a una distribución normal estándar.

Finalmente se menciona (sin demostración) la forma multivariante del comportamiento asintótico de la distribución en el muestreo de $\hat{\boldsymbol{\theta}}(\boldsymbol{z})$,

$$p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) \approx \mathrm{N}_k\{\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, [nI(\boldsymbol{\theta})]^{-1}\},$$

donde $I(\boldsymbol{\theta})$ es la matriz de información de Fisher correspondiente a $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, con elemento general

$$I_{ij}(\boldsymbol{\theta}) = -\int_{\mathcal{X}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(\boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{x}.$$

3.3 Suficiencia y Familia Exponencial

Argumentando una deseable reducción en la dimensionalidad de los datos, en este tema se introduce el concepto de *estadístico suficiente*, definiéndolo directamente como una función de los datos $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{z})$ tal que la función de verosimilitud admite una factorización de la forma

$$l(\boldsymbol{\theta} \mid \boldsymbol{z}) = f(\boldsymbol{t}, \boldsymbol{\theta})\, g(\boldsymbol{z}).$$

Se introduce el concepto de estadístico *minimal suficiente*, se introduce el *estadístico de orden*, y se analizan ejemplos escogidos para estudiar las condiciones de existencia y unicidad de un estadístico minimal suficiente.

Se introduce la *familia exponencial* regular de distribuciones y se comprueba que contiene a la mayor parte de los modelos probabilísticos más comunes. Se demuestra que cualquier elemento de esa familia tiene un estadístico suficiente de dimensión finita e independiente del tamaño muestral y se menciona (sin demostración), que el recíproco también es cierto: las únicas distribuciones con estadístico suficiente de dimensión finita independiente del tamaño muestral son las de la familia exponencial.

Finalmente, se mencionan las extensiones a familias exponenciales no regulares. Para los detalles técnicos, el alumno es referido a la monografía de Huzurbazar (1976).

4 INFERENCIA: MÉTODOS BAYESIANOS OBJETIVOS

La tercera parte del programa describe el proceso de aprendizaje sobre los verdaderos valores de los parámetros, y su uso para predecir el comportamiento de nuevas observaciones, en términos del paradigma bayesiano. Con objeto de obtener resultados directamente utilizables en el contexto de la investigación científica se utilizan exclusivamente distribuciones iniciales objetivas, definidas a partir del modelo probabilístico supuesto mediante el uso de la teoría matemática de la información. En particular, se proporcionan soluciones bayesianas objetivas para los problemas convencionales de *estimación puntual*, de *estimación por intervalos* y de *contraste de hipótesis*, cuyo comportamiento en el muestreo será analizado en la cuarta parte del programa.

4.1 El Paradigma Bayesiano

Formalmente, los métodos bayesianos de inferencia estadística son simplemente una aplicación de la teoría de la probabilidad; resulta asi apropiado que muchos de los textos clásicos de inferencia bayesiana, como los de Laplace (1812), Jeffreys (1939) o de Finetti (1970), lleven como título 'Teoría de la Probabilidad'.

Dado un conjunto de datos $\boldsymbol{z} \in \mathcal{Z}$ cuyo comportamiento probabilístico se asume descrito por un modelo $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, y determinada una distribución inicial para el parámetro $\pi(\boldsymbol{\theta})$, el *teorema de Bayes* (de donde el paradigma toma su nombre) especifica que toda la información disponible sobre el valor de $\boldsymbol{\theta}$ está contenida en su *distribución final*,

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{z}) = \frac{p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int_\Theta p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \ .$$

Los problemas residen en:

(i) la elección de un modelo probabilístico adecuado $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, y

(ii) la especificación de la distribución inicial $\pi(\boldsymbol{\theta})$.

La elección del modelo probabilístico es un problema complejo, cualquiera que sea la postura adoptada ante el problema de inferencia. La elección del modelo típicamente depende tanto de un análisis descriptivo de los datos disponibles como de una interpretación contextual de su comportamiento, y su construcción no será tratada en este programa.

Naturalmente, cualquier resultado inferencial es condicional al modelo supuesto. Se menciona, sin embargo, que el análisis crítico de un modelo específico es posible en el contexto del problema de contraste de hipótesis, descrito en la Sección 4.4. Finalmente se adelanta que, en los métodos bayesianos objetivos, la distribución inicial es una función del modelo probabilístico supuesto, llamada la *distribución inicial de referencia*, y descrita en el tema siguiente.

En el resto de este tema se describe el mecanismo del proceso de aprendizaje bayesiano, dados un modelo $p(\boldsymbol{z} \mid \boldsymbol{\theta})$ y a una distribución inicial $\pi(\boldsymbol{\theta})$. En particular,

(i) se describe el uso de la forma proporcional del teorema de Bayes

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta;$$

(ii) se considera el caso en el que el contexto del problema permite restringir los valores posibles del parámetro a un subconjunto $\Theta_0 \subset \Theta$ del espacio original justificándose que

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{z}, \theta \in \Theta_0) = \begin{cases} \dfrac{\pi(\boldsymbol{\theta} \mid \boldsymbol{z})}{\displaystyle\int_{\Theta_0} \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\theta}, & \text{si } \theta \in \Theta_0 \\ 0, & \text{si } \theta \notin \Theta_0; \end{cases}$$

(iii) se considera el caso en el que el vector $\boldsymbol{\theta}$ se descompone en la forma $\boldsymbol{\theta} = (\phi, \boldsymbol{\omega})$, donde ϕ es el parámetro de interés y $\boldsymbol{\omega}$ un parámetro marginal, y se define la distribución final del parámetro de interés,

$$\pi(\phi \mid \boldsymbol{z}) = \int_{\Omega} \pi(\phi, \boldsymbol{\omega} \mid \boldsymbol{z})\, d\boldsymbol{\omega} \propto \pi(\phi) \int_{\Omega} p(\boldsymbol{z} \mid \phi, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \mid \phi)\, d\boldsymbol{\omega}.$$

Se considera además el caso particular en que los datos $\boldsymbol{z}$ constituyen una muestra aleatoria

$$\boldsymbol{z} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$$

de una distribución $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, de forma que

$$p(\boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_i p(\boldsymbol{x}_i \mid \boldsymbol{\theta}),$$

y se describe la *distribución predictiva* $p(\boldsymbol{x} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de una observación futura, dada por

$$p(\boldsymbol{x} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \int_{\Theta} p(\boldsymbol{x} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\, d\boldsymbol{\theta}.$$

Los conceptos descritos son ilustrados con distintos ejemplos elementales, insistiéndose en el contenido intuitivo de las representaciones gráficas correspondientes. La Figura 3, por ejemplo, describe la distribución final y la distribución predictiva correspondientes a una muestra de 25 observaciones simuladas a partir de una distribución exponencial de parámetro $\theta = 0.5$.

En todos los casos se utilizan como distribuciones iniciales las correspondientes funciones de referencia (por ejemplo $\pi^*(\theta) = \theta^{-1}$ en el caso exponencial), que se justifican en el tema siguiente.

El tema concluye con una descripción del comportamiento asintótico de la distribución final cuando los datos constituyen una muestra aleatoria de gran tamaño. En particular, se esquematiza la demostración de los dos resultados más importantes:

(i) Parámetro discreto $\boldsymbol{\theta}$, con espacio paramétrico numerable, $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots\}$, que incluye al verdadero valor $\boldsymbol{\theta}_t$. En este caso,

$$\lim_{n\to\infty} \Pr[\boldsymbol{\theta} = \boldsymbol{\theta}_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = 1,$$
$$\lim_{n\to\infty} \Pr[\boldsymbol{\theta} = \boldsymbol{\theta}_j \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = 0,\ \forall \boldsymbol{\theta}_j \neq \boldsymbol{\theta}_t.$$

(ii) Parámetro continuo, $\boldsymbol{\theta} \in \Theta \subset \Re^k$. En este caso, para n grande y $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ suficientemente regular,

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \approx N_k(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}, [nI(\hat{\boldsymbol{\theta}})]^{-1}),$$

donde $I(\boldsymbol{\theta})$ es la matriz de Fisher correspondiente a $p(\boldsymbol{x} \mid \boldsymbol{\theta})$.

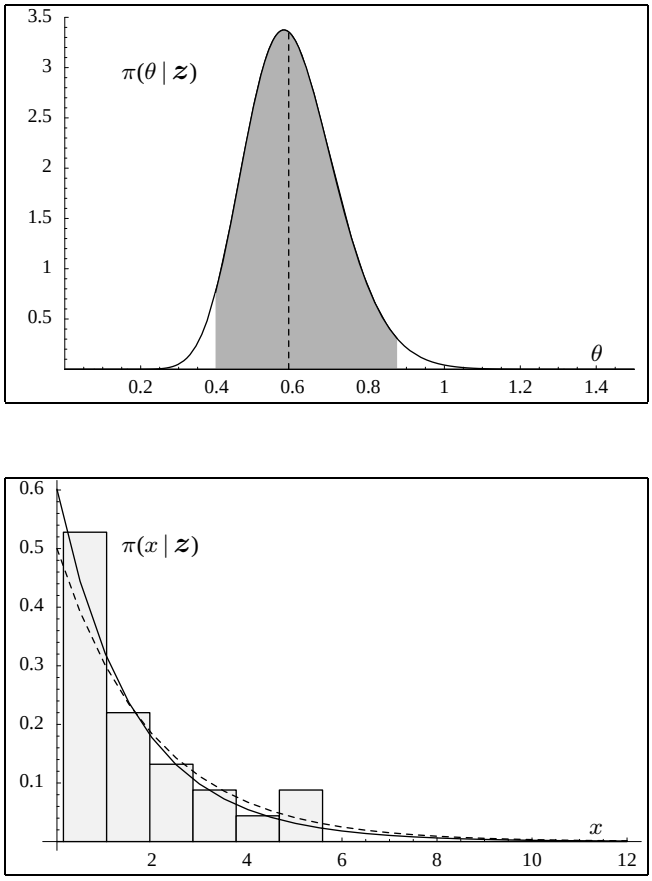

Figura 3. A partir de una muestra aleatoria $\boldsymbol{z} = \{x_1, \ldots, x_n\}$ de 25 observaciones cuya suma es $\sum_i x_i = 41.57$, simuladas de una distribución exponencial con $\theta = 0.5$, se representa (i) la distribución final de θ, con indicación del estimador intrínseco, $\theta^* = 0.59$, y de la región intrínseca 0.95-creíble, $[0.40, 0.88]$, y (ii) la distribución predictiva de una observación futura x, sobreimpuesta a un histograma de los datos y al modelo correcto, una densidad exponencial con parámetro $\theta = 0.5$ (representada con una línea discontinua).

4.2 Distribuciones de Referencia

En este tema se motiva la necesidad de disponer de un 'origen' para el conjunto de distribuciones iniciales, de una función que formalmente actúe como distribución inicial, y que proporcione un contenido matemático preciso a la idea de que no se dispone de información inicial alguna sobre el parámetro de interés.

Se considera en primer lugar el caso univariante, de forma que $\theta \in \Theta \subset \Re$. Se describe como la teoría matemática de la información (Shannon, 1948) *define* la cantidad de información

$$I^{\theta}\{\mathcal{Z}, \pi_\theta(.)\} = \int_{\mathcal{Z}} p(\boldsymbol{z}) \int_{\Theta} \pi(\theta \mid \boldsymbol{z}) \log \frac{\pi(\theta \mid \boldsymbol{z})}{\pi(\theta)}\, d\theta d\boldsymbol{z},$$

que puede esperarse de un conjunto de datos $\boldsymbol{z} \in \mathcal{Z}$ sobre el valor de un parámetro θ. Se trata de una medida

de la información que pueden proporcionar los datos, que depende de la información inicial de que se disponga sobre el valor de θ, descrita por su distribución inicial $\pi_\theta(.)$.

Se pone de manifiesto que tal medida no es más que la discrepancia intrínseca entre la distribución conjunta de $\boldsymbol{z}$ y de θ y el producto de sus distribuciones marginales, es decir su dependencia intrínseca

$$I^\theta\{\mathcal{Z}, \pi_\theta(.)\} = \delta\{p_{z\theta}, p_z\pi_\theta\} = D(p_{z\theta}).$$

Se indica que, si se replicase indefinidamente el experimento original (que consiste en observar $\boldsymbol{z} \in \mathcal{Z}$), se terminaría conociendo el verdadero valor del parámetro θ; consecuentemente, la información proporcionada por k réplicas condicionalmente independientes del mismo experimento, esto es, $I^\theta\{\mathcal{Z}^k, \pi_\theta(.)\}$ convergerá, cuando $k \to \infty$, a la *cantidad de información desconocida* sobre θ cuando la distribución inicial es $\pi(\theta)$.

La *distribución inicial de referencia* $\pi^*(\theta)$ es aquélla que *maximiza la cantidad de información desconocida*, y la distribución final de referencia $\pi^*(\theta \mid \boldsymbol{z})$ es la que resulta de aplicar el teorema de Bayes.

Formalmente, si $\mathcal{P}$ es el conjunto de las distribuciones iniciales de θ compatibles con el problema de inferencia planteado,

$$\begin{aligned}
\pi^*(\theta \mid \boldsymbol{z}) &\propto p(\boldsymbol{z} \mid \theta)\, \pi^*(\theta), \\
\pi^*(\theta \mid \boldsymbol{z}) &= \delta \lim \pi^{(k)}(\theta \mid \boldsymbol{z}), \\
\pi^{(k)}(\theta \mid \boldsymbol{z}) &\propto p(\boldsymbol{z} \mid \theta)\, \pi^{(k)}(\theta), \\
\pi_\theta^{(k)}(\cdot) &= \arg \sup_{\pi_\theta(\cdot)\in\mathcal{P}} I^\theta\{\mathcal{Z}^k, \pi_\theta(\cdot)\}.
\end{aligned}$$

Se enuncian, sin demostración, los dos resultados más importantes del análisis de referencia:

(i) Caso discreto finito. Si $\Theta = \{\theta_1, \ldots, \theta_m\}$,

$$\lim_{k\to\infty} I^\theta\{\mathcal{Z}^k, \pi_\theta(\cdot)\} = -\sum\nolimits_{j=1}^{m} \pi_\theta(\theta_j)\, \log \pi_\theta(\theta_j);$$

consecuentemente, la distribución de referencia

$$\pi^*(\theta) = \arg \sup_{\pi_\theta(.)\in\mathcal{P}} \; -\sum\nolimits_{j=1}^{m} \pi_\theta(\theta_j)\, \log \pi_\theta(\theta_j),$$

es la distribución de *máxima entropía* (uniforme si no existen restricciones adicionales).

(ii) Caso continuo. Si $\Theta \subset \Re$, existe un estimador $\hat\theta$ consistente, asintóticamente suficiente y no hay restricciones adicionales,

$$\pi^*(\theta) \propto \hat\pi(\theta \mid \hat\theta)|_{\hat\theta=\theta},$$

donde $\hat\pi(\theta \mid \hat\theta)$ es una aproximación asintótica *cualquiera* a la distribución final. En particular, si $\hat\pi(\theta \mid \hat\theta)$ es normal, entonces la distribución inicial de referencia es

$$\pi^*(\theta) \propto i(\theta)^{1/2}$$

donde $i(\theta)$ es la función de información de Fisher. Consecuentemente, la distribución de referencia conincide en este caso con distribución inicial de Jeffreys.

El tema concluye describiendo el algoritmo que permite obtener distribuciones de referencia en problemas multivariantes.

En primer lugar, se detalla la forma en la que el problema bivariante, $p(\boldsymbol{z} \mid \phi, \omega)$ puede reducirse a la aplicación sucesiva, en dos etapas, del algoritmo univariante:

(i) condicionando a ϕ (con lo que se trabaja con un único parámetro ω) se determina directamente la distribución de referencia $\pi^*(\omega \mid \phi)$ del parámetro marginal condicionado al parámetro de interés;

(ii) obteniendo el modelo integrado

$$p(\boldsymbol{z} \mid \phi) = \int_\Omega p(\boldsymbol{z} \mid \phi, \omega)\, \pi^*(\omega \mid \phi)\, d\omega,$$

(que ya no depende de ω) se determina la distribución marginal de referencia $\pi^*(\phi)$ del parámetro de interés,

(iii) finalmente, la distribución conjunta de referencia cuando ϕ es el parámetro de interés se define como $\pi_\phi^*(\phi, \omega) = \pi^*(\omega \mid \phi)\, \pi^*(\phi)$.

El procedimiento se generaliza inmediatamente al caso multivariante, con vector paramétrico

$$\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\} \in \Theta \subset \Re^k,$$

mediante un algoritmo secuencial que permite obtener

$$\pi^*(\theta_k \mid \theta_1, \ldots, \theta_{k-1}) \times \cdots \times \pi^*(\theta_2 \mid \theta_1)\pi^*(\theta_1)$$

como la distribución inicial de referencia para la parametrización ordenada $\{\theta_1, \ldots, \theta_k\}$.

La teoría general se ilustra con el ejemplo paradigmático proporcionado por las distribuciones de referencia correspondientes al caso de una distribución normal con ambos parámetros desconocidos. Para una introducción sencilla al análisis bayesiano de referencia, el alumno es referido a Bernardo & Ramón (1998).

4.3 Estimación

El 'estimador' bayesiano de $\boldsymbol{\theta}$ es su distribución final $\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{z})$. Sin embargo, con objeto de facilitar la comprensión del contenido inferencial de $\pi_{\boldsymbol{\theta}}(. \mid \boldsymbol{z})$, es conveniente determinar medidas de localización de $\boldsymbol{\theta}$ (*estimación puntual*) y regiones del espacio paramétrico a las que $\boldsymbol{\theta}$ probablemente pertenece (*estimación por intervalos*).

Entre las numerosas medidas de localización posibles, se destacan, por su facilidad de cálculo y de interpretación, *la media y la moda de la distribución final*; se discuten sus condiciones de existencia y unicidad, y se analiza su robustez frente a pequeñas variaciones en los datos. Se pone de manifiesto sin embargo que tales procedimientos de estimación *no* son invariantes: ni la media ni la moda de una transformación biyectiva del parámetro $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ coinciden, en general, con los valores transformados de la media y la moda de $\boldsymbol{\theta}$. Se destaca que, en problemas univariantes, la *mediana de la distribución final* resulta preferible (es invariante y es más robusta frente a pequeñas variaciones en los datos), pero no resulta fácilmente generalizable a problemas multivariantes.

Con respecto a la estimación por intervalos se define *una región de confianza p-creíble* $C_p^{\theta}(\boldsymbol{z})$ como un subconjunto cualquiera del espacio paramétrico con probabilidad final p, de forma que

$$C_p^{\theta}(\boldsymbol{z}) \subset \Theta, \quad \int_{C_p^{\theta}(\boldsymbol{z})} \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta} = p.$$

Se destaca que se trata de un concepto invariante: para cualquier transformación biyectiva del parámetro $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, $C_p^{\boldsymbol{\phi}} = \boldsymbol{\phi}(C_p^{\theta})$ es una región de confianza p-creíble para $\boldsymbol{\phi}$.

Se pone de manifiesto que, en general, existen infinitas regiones p-creíbles y se mencionan, por su facilidad de interpretación, las regiones de máxima densidad final de probabilidad, comentándose que son frecuentemente únicas para una determinada parametrización, pero que la propiedad de máxima densidad de probabilidad final *no* es invariante ante transformaciones biyectivas del parámetro.

El tema concluye con una descripción del problema de estimación en términos de la teoría de la decisión, lo que permite definir la solución óptima correspondiente a un sistema de preferencias determinado.

Formalmente, un problema de estimación puntual, es un *problema de decisión* en el que el conjunto de alternativas coincide con el espacio paramétrico. Para una función de pérdida $L(\boldsymbol{\theta}^e, \boldsymbol{\theta})$ que describe las consecuencias de utilizar un estimador $\boldsymbol{\theta}^e$ cuando el verdadero valor del parámetro es $\boldsymbol{\theta}$, el *estimador Bayes* $\boldsymbol{\theta}^b(\boldsymbol{z})$ es aquél que minimiza la pérdida esperada; formalmente,

$$\boldsymbol{\theta}^b(\boldsymbol{z}) = \arg\min_{\boldsymbol{\theta}^e \in \Theta} \int_{\Theta} L(\boldsymbol{\theta}^e, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta}.$$

Se demuestra que la media y la moda finales son las soluciones que respectivamente corresponden a las funciones de pérdida cuadrática y dicotómica límite y que, en el caso univariante, la mediana corresponde a una función lineal a trozos y simétrica.

Se pone de manifiesto que el uso de funciones de pérdida invariantes garantiza la obtención de estimadores invariantes. En particular, se define el *estimador intrínseco* $\boldsymbol{\theta}^*(\boldsymbol{z})$ como el estimador Bayes correspondiente a la función de pérdida intrínseca (cf. Sección 2.3), esto es el que minimiza la *pérdida intrínseca esperada* con respecto a la distribución de referencia,

$$d(\boldsymbol{\theta}^e \mid \boldsymbol{z}) = \int_{\Theta} \delta\{\boldsymbol{\theta}^e, \boldsymbol{\theta}\}\, \pi^*(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta},$$

donde

$$\delta\{\boldsymbol{\theta}^e, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{z}}(. \mid \boldsymbol{\theta}^e), p_{\boldsymbol{z}}(. \mid \boldsymbol{\theta})\}$$

denota la discrepancia intrínseca entre los dos modelos $p(\boldsymbol{z} \mid \boldsymbol{\theta}^e)$ y $p(\boldsymbol{z} \mid \boldsymbol{\theta})$, respectivamente identificados por $\boldsymbol{\theta}^e$ y por $\boldsymbol{\theta}$.

El tema concluye describiendo la construcción de regiones creíbles invariantes (generalmente únicas) a partir de la pérdida intrínseca esperada. En efecto, para cualquier conjunto de datos $\boldsymbol{z} \in \mathcal{Z}$, el conjunto $C_d(\boldsymbol{z}) \subset \Theta$ de valores de $\boldsymbol{\theta}^e$ para los que la pérdida intrínseca esperada es menor que una determinada constante d, esto es,

$$C_d(\boldsymbol{z}) = \{\boldsymbol{\theta}^e \in \Theta;\ d(\boldsymbol{\theta}^e \mid \boldsymbol{z}) \leq d\}$$

constituye una región invariante ante transformaciones biyectivas de $\boldsymbol{\theta}$ y, bajo condiciones muy generales, la constante d puede ser elegida con la condición de que la probabilidad final de que $\boldsymbol{\theta}$ pertenezca a $C_d(\boldsymbol{z})$ sea un valor cualquiera $p \in (0, 1)$.

Formalmente, una *región intrínseca p-creíble* es un subconjunto $C_p^*(\boldsymbol{z}) \subset \Theta$ del espacio paramétrico tal que

$$\int_{C_p^*(\boldsymbol{z})} \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta} = p,$$

y para todo $\boldsymbol{\theta}_i \notin C_p^*(\boldsymbol{z})$ existe un $\boldsymbol{\theta}_j \in C_p^*(\boldsymbol{z})$ tal que

$$d(\boldsymbol{\theta}_j \mid \boldsymbol{z}) < d(\boldsymbol{\theta}_i \mid \boldsymbol{z}).$$

En la Figura 4 se representa la función de pérdida intrínseca esperada $d(\boldsymbol{\theta}^e \mid \boldsymbol{z})$ correspondiente a una muestra aleatoria de 25 observaciones exponenciales, y se indican tanto el estimador intrínseco como la región intrínseca 0.95-creíble. Para los detalles técnicos referentes a la estimación intrínseca, el alumno es referido a Bernardo & Juárez (2003).

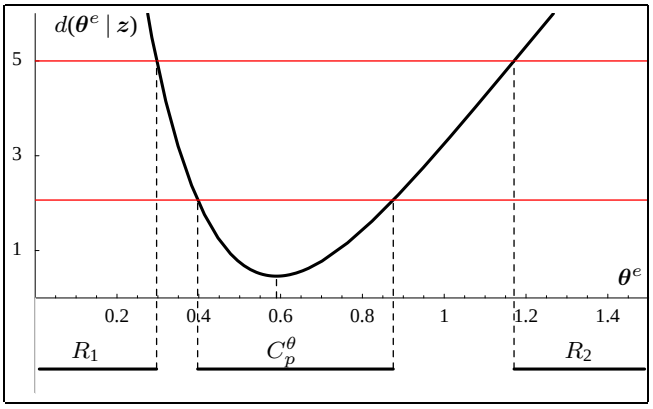

Figura 4. Función de pérdida esperada intrínseca correspondiente a los datos exponenciales analizados en la Figura 3. Se indican el mínimo (correspondiente al estimador intrínseco), el intervalo intrínseco 0.95-creíble C_p^θ (correspondiente a $d(\boldsymbol{\theta}^e \mid \boldsymbol{z}) \leq 2.07$) y las regiones de rechazo R_1 y R_2 correspondientes al contraste definido por $d(\boldsymbol{\theta}^e \mid \boldsymbol{z}) > 5$.

4.4 Contraste de Hipótesis

Dado un modelo probabilístico $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, que hipotéticamente describe el comportamiento del vector observable $\boldsymbol{z} \in \mathcal{Z}$ en función de un vector paramétrico desconocido $\boldsymbol{\theta} \in \Theta$, se trata de decidir si los datos observados $\boldsymbol{z}$ son o no son *compatibles* con la hipótesis $H_0 \equiv \{\boldsymbol{\theta} \in \Theta_0\}$ de que $\boldsymbol{\theta}$ pertenece a un determinado subconjunto $\Theta_0 \subset \Theta$ del espacio paramétrico.

Formalmente, se trata de un problema de decisión con sólo dos alternativas, aceptar (a_0) o rechazar (a_1) la hipótesis de trabajo o *hipótesis nula* H_0, cuya solución normativa tan sólo depende de la *pérdida diferencial*

$$\Delta L(\boldsymbol{\theta}, \Theta_0) = L(a_0, \boldsymbol{\theta}) - L(a_1, \boldsymbol{\theta})$$

que mide (en función de $\boldsymbol{\theta}$) el incremento de pérdida que podría originar la aceptación de la hipótesis nula. Se observa que a_0 constituye una simplificación del modelo y que, consecuentemente, debe existir alguna ventaja en aceptar H_0 cuando es cierta, de forma que, necesariamente,

$$\Delta L(\boldsymbol{\theta}, \Theta_0) < 0, \ \forall \boldsymbol{\theta} \in \Theta_0;$$

si H_0 es cierta, la pérdida diferencial por aceptarla es negativa.

Se pone de manifiesto que una función de pérdida dicotómica, del tipo $\Delta L(\boldsymbol{\theta}, \Theta_0) = \pm 1$ según $\boldsymbol{\theta}$ pertenezca o no a Θ_0, da lugar a un criterio probabilístico:

$$\text{Rechazar } H_0 \iff \Pr(\boldsymbol{\theta} \in \Theta_0 \mid \boldsymbol{z}) < \Pr(\boldsymbol{\theta} \notin \Theta_0 \mid \boldsymbol{z}).$$

En el caso particular en el que la probabilidad inicial de H_0 sea $1/2$, el criterio se reduce a determinar el *factor de Bayes*,

$$B_{01}(\boldsymbol{z}) = \frac{\int_{\Theta_0} p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid H_0)\, d\boldsymbol{\theta}}{\int_{\Theta_1} p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid H_1)\, d\boldsymbol{\theta}}$$

donde $\Theta_1 = \Theta - \Theta_0$, y donde $\pi(\boldsymbol{\theta} \mid H_0)$ y $\pi(\boldsymbol{\theta} \mid H_1)$ son respectivamente las distribuciones iniciales bajo la hipótesis nula y bajo la hipótoesis alternativa, que no son necesariamente especificadas como restricciones de una distribución común $\pi(\boldsymbol{\theta})$. En estas condiciones, se rechaza la hipótesis de trabajo si (y sólamente si) $B_{01}(\boldsymbol{z}) < 1$. Cuando sólo existen dos valores posibles de $\boldsymbol{\theta}$ (hipótesis simples), de forma que $\Theta = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$, el factor de Bayes se reduce al cociente de verosimilitudes.

Se observa que en el caso de hipótesis *precisas*, cuando Θ_0 es un conjunto de medida nula (por ejemplo, cuando $\boldsymbol{\theta}$ es un parámetro continuo y Θ_0 contiene un único valor $\boldsymbol{\theta}_0$) esta formulación exige una distribución inicial *mixta*, que asigne una masa de probabilidad a un punto. Tal planteamiento responde por lo tanto a situaciones caracterizadas por importante información previa (probabilidad inicial muy concentrada alrededor de $\boldsymbol{\theta}_0$), y permite valorar hasta que punto tal información inicial es modificada por los datos. Se trata pues de un problema muy distinto al problema planteado, que es el de determinar la *compatibilidad* de los datos observados con el valor $\boldsymbol{\theta}_0$ *en ausencia de información inicial* sobre el verdadero valor de $\boldsymbol{\theta}$.

Para obtener una solución general al problema de decidir si los datos son o no son compatibles con una hipótesis de la forma $\boldsymbol{\theta} \in \Theta_0$, donde $\boldsymbol{\theta}$ es un parámetro continuo, es necesario utilizar una función diferencial de pérdida $\Delta L(\boldsymbol{\theta}, \Theta_0)$ que sea continua en $\boldsymbol{\theta}$. La solución normativa es entonces simplemente

$$\text{Rechazar } H_0 \iff \int_{\Theta} \Delta L(\boldsymbol{\theta}, \Theta_0)\, \pi(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta} > 0,$$

que únicamente requiere la determinación de la distribución final $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ (generalmente construida a partir de la distribución de referencia $\pi^*(\boldsymbol{\theta})$, sin información inicial sobre el valor de $\boldsymbol{\theta}$).

Se pone de manifiesto que la discrepancia intrínseca proporciona una solución natural al problema de especificar la función de pérdida diferencial, solución que resulta además invariante frente a transformaciones biyectivas del parámetro. Formalmente, se propone el uso de la función de pérdida diferencial

$$\Delta L(\boldsymbol{\theta}, \Theta_0) = \inf_{\boldsymbol{\theta}_0 \in \Theta_0} \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}) - \delta^*, \quad \delta^* > 0,$$

donde δ^* representa la utilidad esperada de aceptar la hipóteis nula cuando es cierta, y donde

$$\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{z}}(. \mid \boldsymbol{\theta}_0), p_{\boldsymbol{z}}(. \mid \boldsymbol{\theta})\}$$

denota la discrepancia intrínseca entre los dos modelos $p(\boldsymbol{z} \mid \boldsymbol{\theta}_0)$ y $p(\boldsymbol{z} \mid \boldsymbol{\theta})$.

La correspondiente solución normativa, (el *criterio bayesiano de referencia* o BRC, Bernardo, 1999; Bernardo & Rueda, 2002), resulta ser

$$\text{Rechazar} \quad H_0 \iff d(\Theta_0 \mid \boldsymbol{z}) > \delta^*,$$

$$d(\Theta_0 \mid \boldsymbol{z}) = \int_{\Theta} \left\{ \inf_{\boldsymbol{\theta}_0 \in \Theta_0} \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \right\} \pi^*(\boldsymbol{\theta} \mid \boldsymbol{z})\, d\boldsymbol{\theta} > 0,$$

donde

$$\pi^*(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto p(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi^*(\boldsymbol{\theta})$$

es la distribución final de referencia.

Se observa que, como función de $\boldsymbol{z} \in \mathcal{Z}$, el procedimiento produce de forma *automática* una función de contraste (positiva e invariante), el *estadístico bayesiano de referencia*,

$$b_0(\boldsymbol{z}) = d(\Theta_0 \mid \boldsymbol{z}), \quad \boldsymbol{z} \in \mathcal{Z},$$

y que el criterio consiste en rechazar la hipótesis de trabajo si, y sólamente si, el valor de $b_0(\boldsymbol{z})$ es suficientemente alto.

Se demuestra que el valor esperado de $b_0(\boldsymbol{z})$ bajo la hipótesis nula converge a la unidad para muestras grandes (y en muchos casos es exactamente igual a uno para cualquier tamaño muestral).

Además, puesto que la discrepancia intrínseca entre dos modelos probabilísticos es el mínimo valor esperado del logaritmo del cociente de verosimilitudes en favor del modelo verdadero (cf. Sección 2.3), valores de $b_0(\boldsymbol{z})$ superiores a 2.5 suponen cocientes de verosimilitudes esperados del orden de $e^{2.5} \approx 12$, mientras que valores de $b_0(\boldsymbol{z})$ superiores a 5.0 suponen cocientes de verosimilitudes esperados del orden de $e^{5.0} \approx 150$.

Consecuentemente, para *cualquier tamaño muestral y para cualquier dimensionalidad* del problema, el criterio bayesiano de referencia sugiere aceptar hipótesis con discrepancias intrínsecas esperadas cercanas a la unidad, buscar más información con valores situados alrededor de 2.5, y rechazar hipótesis con discrepancias intrínsecas esperadas mayores que 5.

En la Figura 4 se representan las regiones de rechazo correspondientes a $b_0(\boldsymbol{z}) > 5$ relativas a los datos exponenciales descritos y analizados en la Figura 3.

En el problema paradigmático del contraste de hipótesis sobre una media normal, con $H_0 \equiv \{\mu = \mu_0\}$, los valores 2.5 y 5.0 corresponden al caso en el que μ_0 se sitúa, respectivamente, a 2 y 3 desviaciones típicas de $\bar{x}$.

5 CALIBRACIÓN: MÉTODOS FRECUENTISTAS

En la tercera parte del programa se han descrito un conjunto de procedimientos inferenciales, construidos desde una perspectiva bayesiana a partir de distribuciones iniciales objetivas. En la cuarta parte del programa se analiza el comportamiento en el muestreo de tales procedimientos y se compara el resultado con el comportamiento de otros procedimientos descritos en la literatura. En particular, se hace uso de los métodos frecuentistas para *calibrar* los métodos bayesianos objetivos, estudiando su comportamiento bajo muestreo repetido.

El contenido de esta cuarta parte está estructurado en cuatro temas. En el primero de ellos se analizan las aportaciones que pueden esperarse del estudio del comportamiento en el muestreo de un procedimiento estadístico, se mencionan algunas de sus limitaciones, y se ilustra la variabilidad en el muestreo de la propia distribución final. Los otros tres temas están respectivamente dedicados a estudiar el comportamiento bajo muestreo repetido de estimadores puntuales, regiones de confianza y procedimientos de contraste de hipótesis.

5.1 Comportamiento Medio de un Procedimiento Estadístico

Para datos $\boldsymbol{z} \in \mathcal{Z}$, cuyo comportamiento probabilístico se asume descrito por $\{p(\boldsymbol{z} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, se considera un *procedimiento* estadístico definido por una transformación de los datos $\boldsymbol{t}(\boldsymbol{z})$.

Ejemplos específicos incluyen un estimador puntual, una región de confianza, un estadístico de contraste, y la propia distribución final $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$.

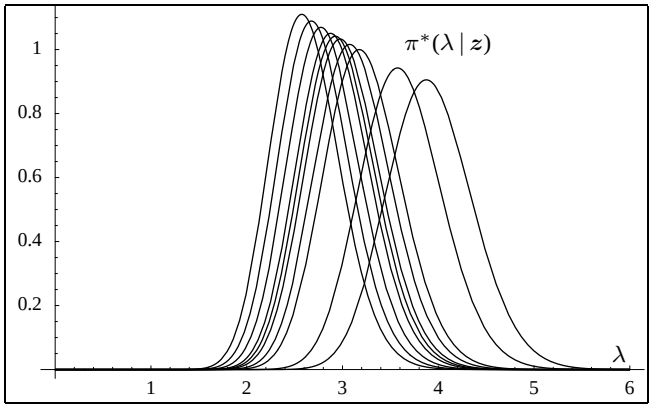

Figura 5. Distribuciones finales de referencia para el parámetro λ de una distribución de Poisson correspondientes a 10 muestras aleatorias de tamaño $n = 20$ simuladas a partir de una distribución de Poisson con parámetro $\lambda = 3$.

En la Figura 5 se representan las distribuciones finales de referencia correspondientes a 10 muestras aleatorias de tamaño 20 simuladas a partir de una distribución de Poisson de parámetro $\lambda = 3$, lo que ilustra la variabilidad de las conclusiones inferenciales en función de los datos observados.

Para un procedimiento concreto $\boldsymbol{t}(\boldsymbol{z})$ es posible definir una *función de pérdida* $L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}$ que, de alguna forma, debe medir el 'error' cometido por el procedimiento en función del verdadero valor del vector paramétrico $\boldsymbol{\theta}$.

Considerada como función de $\boldsymbol{z}$ (condicional a $\boldsymbol{\theta}$), $L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}$ es una variable aleatoria cuya distribución en el muestreo puede ser determinada, en principio, a partir de la distribución condicional de $\boldsymbol{z}$, $p(\boldsymbol{z} \mid \boldsymbol{\theta})$. Consecuentemente, su valor esperado,

$$\begin{aligned} r_t(\boldsymbol{\theta} \mid L) &= \mathrm{E}_{\boldsymbol{z} \mid \boldsymbol{\theta}}[L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}] \\ &= \int_{\mathcal{Z}} L\{\boldsymbol{t}(\boldsymbol{z}), \boldsymbol{\theta}\}\, p(\boldsymbol{z} \mid \boldsymbol{\theta})\, d\boldsymbol{z}, \end{aligned}$$

puede ser interpretado como el *riesgo medio* (en muestreo repetido) que se corre utilizando ese procedimiento.

Como ejemplos de funciones de pérdida habituales se mencionan la pérdida cuadrática asociada a un estimador puntual $\tilde{\boldsymbol{\theta}}$,

$$L\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

y la pérdida logarítmica (la función de evaluación cambiada de signo), asociada a una distribución final $\pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{z})$,

$$L\{\pi_{\boldsymbol{\theta}}(. \mid \boldsymbol{z}), \boldsymbol{\theta}\} = -\log[\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta} \mid \boldsymbol{z})].$$

Para que un procedimiento estadístico resulte aceptable es naturalmente *necesario* que tenga un riesgo medio pequeño frente a funciones de pérdida razonables. Es obvio, sin embargo, que ésta *no* es una condición suficiente: un buen comportamiento medio *no garantiza* un buen comportamiento en cada caso concreto. Como se demuestra con ejemplos, un procedimiento puede tener un comportamiento claramente inaceptable en casos concretos, pero presentar un buen comportamiento medio debido a que tales casos no son muy frecuentes; es posible incluso encontrar procedimientos que tienen un mal comportamiento en *todos* los casos pero que presentan un buen comportamiento medio, debido a una compensación de los errores.

En general el riesgo medio de utilizar un procedimiento $\boldsymbol{t}$, depende del verdadero valor del parámetro $\boldsymbol{\theta}$. Es sin embargo posible que (con respecto a una función de pérdida L), un procedimiento $\boldsymbol{t}_1$ sea *uniformemente* mejor que otro procedimiento $\boldsymbol{t}_2$ en el sentido de que,

$$\forall \boldsymbol{\theta} \in \Theta, \quad r_{\boldsymbol{t}_1}(\boldsymbol{\theta} \mid L) < r_{\boldsymbol{t}_2}(\boldsymbol{\theta} \mid L);$$

en este caso se dice que $\boldsymbol{t}_2$ está *dominado* por $\boldsymbol{t}_1$.

Un procedimiento es declarado *inadmisible* (con respecto a una función de pérdida) si es posible encontrar otro procedimiento que lo domina. Con estos elementos se enuncia (sin demostración) un resultado importante (Savage, 1954): bajo condiciones muy generales, para que un procedimiento sea admisible es *necesario y suficiente* que sea un procedimiento Bayes, esto es que pueda ser obtenido minimizando alguna pérdida esperada. Finalmente se subraya el hecho de que cualquier resultado de admisibilidad es *condicional* a la función de pérdida elegida: un procedimiento inadmisible bajo una función de pérdida determinada puede resultar admisible bajo una función de pérdida diferente.

5.2 Error Medio de los Estimadores Puntuales

El tema empieza introduciendo algunos de los métodos de estimación no bayesiana propuestos en la literatura: se recuerda la estimación máximo verosímil, y se describen los métodos de estimación por momentos, por mínimos cuadrados y mediante estadísticos de orden. Se determinan los estimadores correspondientes para un conjunto de ejemplos representativos, comparándolos con los estimadores Bayes descritos en la Sección 3.3. Se recuerda el concepto de estadístico suficiente y se pone de manifiesto que, frecuentemente, los estimadores puntuales son estadísticos suficientes.

Se argumenta la importancia de un comportamiento límite adecuado. Se introduce el concepto de *consistencia*, y se demuestra que todo estimador Bayes es consistente. Se recuerda el comportamiento asintótico de la distribución en el muestreo del estimador máximo verosímil, subrayándose que, bajo condiciones de regularidad, permite obtener una buena aproximación, para muestras grandes, a la distribución en el muestreo del estimador bayesiano intrínseco.

Se describe la función de pérdida cuadrática como criterio de evaluación del comportamiento de un estimador, se menciona la descomposición del riesgo cuadrático medio en varianza más sesgo al cuadrado y, en este contexto, se introduce el concepto de *eficiencia*. Se menciona el concepto de estimador insesgado y se ilustran, mediante ejemplos, las graves consecuencias a las que puede llevar la insistencia en el uso de estimadores insesgados. Mediante datos simulados se hace un análisis comparativo del comportamiento en el muestreo de los estimadores considerados con respecto al riesgo cuadrático medio.

El tema concluye argumentando las ventajas de la función de pérdida intrínseca como criterio de evaluación frente a la convencional pérdida cuadrática. Se determina una aproximación asintótica para el riesgo intrínseco medio, y se retoman los ejemplos anteriores para proporcionar un análisis comparativo del comportamiento en el muestreo de los distintos estimadores con respecto al riesgo intrínseco medio.

5.3 Recubrimiento Esperado de las Regiones de Confianza

El tema comienza con la descripción del método convencional de construcción de regiones frecuencialistas de confianza a partir de *pivotes* (funciones de los datos y del parámetro de interés cuya distribución en el muestreo es totalmente conocida), destacando la utilidad, en este contexto, de la función de soporte normalizada. Se pone de manifiesto que, mientras que las regiones de confianza p-creíbles (descritas en la Sección 4.3) pueden ser construidas para cualquier problema, los métodos frecuentistas no siempre resultan aplicables.

Se propone la calibración de regiones de confianza p-creíbles en términos de la proporción de muestras $\hat{p}(\theta)$ para las que la correspondiente región p-creíble cubrirá (en muestreo repetido) el verdadero valor del parámetro θ.

Centrándose en el caso de modelos de parámetro unidimensional continuo, se define un intervalo p-creíble unilateral $]-\infty, \theta_p(\boldsymbol{z})]$ mediante el cuantil correspondiente $\theta_p(\boldsymbol{z})$ de la distribución final

$$\Pr[\theta \leq \theta_p(\boldsymbol{z}) \mid \boldsymbol{z}] = \int_{\{\theta \leq \theta_p(\boldsymbol{z})\}} \pi(\theta \mid \boldsymbol{z})\, d\theta = p.$$

Se define el nivel de recubrimiento correspondiente,

$$\Pr[\theta_p(\boldsymbol{z}) \geq \theta \mid \theta] = \int_{\{\theta_p(\boldsymbol{z}) \geq \theta\}} p(\boldsymbol{z} \mid \theta)\, d\boldsymbol{z} = \hat{p}(\theta),$$

donde $p(\boldsymbol{z} \mid \theta)$ es el modelo probabilístico supuesto. Se pone de manifiesto que, para la distribución final de referencia,

$$\hat{p}(\theta) = p + O(n^{-1})$$

de forma que, bajo condiciones de regularidad, una región de confianza p-creíble es siempre, para muestras suficientemente grandes, una región frecuentista de confianza *aproximada*, y que esta aproximación mejora cuando se utilizan distribuciones iniciales de referencia.

Se indica que cuando existe un pivote, (como en los casos de datos normales o de datos exponenciales) se obtiene una doble interpretación *exacta* de las regiones de confianza p-creíbles como regiones frecuentistas con probabilidad de recubrimiento p, pero se subraya que, en general, se trata sólo de una *aproximación asintótica*, con problemas en el caso de muestras extremas (lo que se ilustra con datos binomiales).

Mediante el uso de ejemplos apropiados se pone de manifiesto que, en general, *no es deseable* una coincidencia *exacta* entre p y $\hat{p}(\theta)$. En efecto, los intervalos de confianza frecuentistas, que por construcción garantizan un recubrimiento predeterminado p, pueden dar lugar a resultados indeseables: se trata de otro ejemplo de un valor medio correcto basado en la compensación de errores individuales.

5.4 Probabilidades de Error en los Contrastes de Hipótesis

Se describe el concepto convencional de contraste de significación ('significance test') en el que, sin plantear alternativas, se trata de contrastar una hipótesis puntual $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ proponiendo un estadístico de contraste $t(\boldsymbol{z}) \in \mathcal{T}$ y una región crítica $R \subset \mathcal{T}$, de forma que la probabilidad de rechazar la hipótesis de trabajo (o *hipótesis nula*) cuando es cierta, esto es $\Pr(\boldsymbol{t} \in R \mid H_0)$ o *error de tipo I*, sea una constante α, llamada *nivel de significación*. Se pone de manifiesto que el procedimiento no incluye la forma de elegir α, y que el uso sistemático de valores convencionales para el nivel de significación (como el ubicuo $\alpha = 0.05$) para cualquier tamaño muestral o dimensionalidad da lugar a contradicciones.

Se subraya que para poder elegir de forma razonable entre los infinitos contrastes con el mismo nivel de significación es necesario considerar además la probabilidad de aceptar una hipótesis falsa, $\beta(\boldsymbol{\theta}) = \Pr(\boldsymbol{t} \notin R \mid \boldsymbol{\theta})$ o *error de tipo II*, que depende del verdadero (y desconocido) valor del parámetro $\boldsymbol{\theta}$, y se pone de manifiesto que su cálculo *requiere* especificar una alternativa H_1 a la hipótesis nula.

En este contexto, se pone de manifiesto que el *procedimiento* convencionalmente sugerido y (lamentablemente) utilizado con mucha frecuencia consiste en utilizar aquel contraste que, para una probabilidad α de error de tipo I predeterminada, minimice la probabilidad $\beta(\boldsymbol{\theta})$ de error de tipo II (si fuese posible, para todo valor de $\boldsymbol{\theta}$). Se demuestra, sin embargo, que este *criterio* es *incompatible* con los axiomas de comportamiento coherente que dictan (como la intuición sugiere) que lo que se debe minimizar es una combinación lineal de los dos tipos de error,

$$L(R, \boldsymbol{\theta}, \gamma) = \gamma \Pr(\boldsymbol{t} \in R \mid \boldsymbol{\theta}_0) + (1-\gamma)\Pr(\boldsymbol{t} \notin R \mid \boldsymbol{\theta}),$$

donde los pesos relativos, $\{\gamma, (1-\gamma)\}$, con $0 < \gamma < 1$, miden las pérdidas relativas correspondientes a cada tipo de error, y dependen del contexto del problema.

Se define la *potencia* del contraste como la probabilidad de rechazar la hipótesis nula en función del verdadero valor del parámetro,

$$\text{pot}(\boldsymbol{\theta}) = \Pr(\boldsymbol{t} \in R \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

y se menciona su utilidad para comparar entre sí las características operativas de contrastes alternativos. Cuando la hipótesis nula es una hipótesis simple $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$, de forma que Θ_0 es un conjunto de medida nula, $\text{pot}(\boldsymbol{\theta}) = 1 - \beta(\boldsymbol{\theta})$ es esencialmente la probabilidad de rechazar una hipótesis falsa. Se describe el cociente de verosimilitudes como una forma intuitiva de construir un estadístico de contraste y se enuncia el lema de Neyman-Pearson que garantiza, en el caso de hipótesis simples, la obtención de un contraste de máxima potencia entre los que tienen un nivel de significación dado.

El planteamiento frecuentista de contraste de hipótesis se compara con el planteamiento bayesiano descrito en la Sección 4.4. Se señala que en ambos casos se trata de construir un estadístico de contraste y de rechazar la hipótesis de trabajo cuando el valor numérico de tal estadístico es suficientemente alto. Se hace observar, sin embargo, una diferencia radical: en el caso bayesiano el punto de corte forma parte de la función de pérdida y describe, en unidades de información, la discrepancia que se está dispuesto a tolerar entre el modelo verdadero y el modelo especificado por la hipótesis nula; en el caso frecuentista el punto de corte viene definido en términos de la distribución en el muestreo bajo la hipótesis nula del estadístico de contraste que se escoja lo que, como se ha mencionado, puede dar lugar a contradicciones.

El tema concluye estudiando el comportamiento bajo muestreo repetido de distintos estadísticos de contraste, con especial énfasis en los que definen el cociente de verosimilitudes y la discrepancia intrínseca esperada. En particular, se estudian las probabilidades de error (en muestreo repetido) que corresponden a tales criterios en algunos ejemplos representativos, y se determinan y comparan las funciones de potencia correspondientes.

6 APLICACIONES

A lo largo del programa se habrán estudiado, en forma de ejemplos y de problemas resueltos, muchos problemas de inferencia planteados en modelos sencillos univariantes: inferencia y predicción con datos discretos (Binomiales y Poisson), e inferencia y predicción con datos continuos dependientes de un solo parámetro, tanto en casos regulares (datos exponenciales) como no regulares (datos uniformes). Se habrán estudiado asimismo algunos problemas sencillos de inferencia y predicción con datos continuos dependientes de varios parámetros (datos normales y datos uniformes con ambos extremos desconocidos). En esta última parte del programa se utilizan los conocimientos adquiridos para estudiar problemas algo más complejos.

6.1 Comparación de Proporciones

Se parte de muestras aleatorias $\boldsymbol{x}$, $\boldsymbol{y}$ de observaciones Bernoulli procedentes de poblaciones distintas, con estadísticos suficientes (r_1, n_1) y (r_2, n_2) y con parámetros θ_1 y θ_2, y se supone que interesa comparar los valores de θ_1 y θ_2, proporcionando al alumno ejemplos reales en diferentes contextos (elecciones, marketing, biotecnología, ...).

Se determinan las distribuciones finales de referencia $\pi^*(\phi_i \mid \boldsymbol{x}, \boldsymbol{y})$, para dos posibles funciones de interés, $\phi_1 = \theta_1 - \theta_2$ y $\phi_2 = \theta_1/\theta_2$, se obtienen estimadores puntuales y regiones de confianza para ambas funciones, y se formulan y analizan distintos problemas de contraste de hipótesis. Los resultados teóricos obtenidos son evaluados por simulación y posteriormente aplicados a un ejemplo con datos reales.

6.2 Inferencia con Medias Normales

Se parte de muestras aleatorias $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ de observaciones procedentes de dos poblaciones normales, $\text{N}(x \mid \mu_1, \sigma_1^2)$, $\text{N}(y \mid \mu_2, \sigma_2^2)$, con estadísticos suficientes $(n, \overline{x}, s_x^2)$ y $(m, \overline{y}, s_y^2)$, y se supone que interesa comparar los valores de μ_1 y μ_2, proporcionando al alumno ejemplos reales en diferentes contextos (producción industrial, consumo, ...).

Se determinan las distribuciones finales de referencia $\pi^*(\phi_i \mid \boldsymbol{x}, \boldsymbol{y})$ de dos posibles funciones de interés, $\phi_1 = \mu_1 - \mu_2$ (problema de Behrens-Fisher) y $\phi_2 = \mu_1/\mu_2$, (problema de Fieller-Creasy), se obtienen estimadores puntuales y regiones de confianza para ambas funciones, y se formulan y resuelven distintos problemas de contraste de hipótesis. De nuevo, los resultados obtenidos son evaluados por simulación y aplicados a un ejemplo con datos reales.

Partiendo de los mismos datos $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, e $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ se plantea el problema de hacer inferencia sobre el producto de las medias $\theta = \mu_1\,\mu_2$, motivándolo con ejemplos reales (cálculo de superficies, teoría de errores en física o ingeniería, ...).

Se determina la correspondiente distribución final de referencia $\pi^*(\theta \mid \boldsymbol{x}, \boldsymbol{y})$, se obtienen estimadores puntuales y regiones de confianza para θ, y se formulan distintos contrastes de hipótesis. Una vez más, los resultados son evaluados por simulación y aplicados a un ejemplo con datos reales. Finalmente, se describe (sin demostración) la solución al problema general de hacer inferencia sobre el producto $\theta = \prod_{i=1}^{k} \mu_i$ de k medias normales.

6.3 Regresión

Se parte de un conjunto de datos apareareados $\boldsymbol{z} = \{(y_1, \boldsymbol{x}_1) \ldots, (y_n, \boldsymbol{x}_n)\}$, donde $y_j \in \Re$ es la respuesta observada en condiciones $\boldsymbol{x}_j \in \Re^k$, y se supone que existe una relación funcional aproximada entre y_j y $\boldsymbol{x}_j$, de forma que $y_j = f(\boldsymbol{x}_j \mid \boldsymbol{\theta}) + \epsilon_j$, con $\mathrm{E}[\epsilon] = 0$, y $\mathrm{Var}[\epsilon] = \sigma^2$. Se analiza la representación cartesiana de este tipo de datos en el caso $k = 1$, motivando el problema con distintos conjuntos de datos reales.

Especificado un modelo $\pi(\epsilon \mid \sigma)$ que describa el comportamiento probabilístico de los errores, se formula el problema general de predicción asociado a esta estructura de datos, que requiere el cálculo de la distribución predictiva condicional

$$p(y \mid \boldsymbol{x}, \boldsymbol{z}) = \int_{\Theta} \int_{0}^{\infty} p(y \mid \boldsymbol{x}, \boldsymbol{\theta}, \sigma)\, \pi^*(\boldsymbol{\theta}, \sigma \mid \boldsymbol{z})\, d\boldsymbol{\theta} d\sigma.$$

Bajo normalidad homocedástica, $N(\epsilon \mid 0, \sigma^2)$, se detalla la solución analítica para el caso lineal univariante

$$f(x \mid \theta_0, \theta_1) = \theta_0 + \theta_1\, x,$$

determinándose las distribuciones finales (marginales) de θ_0 y θ_1, la distribución final de σ y la distribución predictiva $p(y \mid x, \boldsymbol{z})$. Se obtienen estimadores puntuales y regiones creíbles para y dado x, y se formulan problemas de contraste para los valores de las θ_i. Los resultados obtenidos son evaluados por simulación y aplicados a un ejemplo con datos reales.

Se enuncia (sin demostración) la generalización de algunos de los resultados anteriores al caso lineal multivariante

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \theta_0 + \sum_{i=1}^{k} \theta_i\, x_i,$$

y se muestra la aplicación de este resultado a problemas univariantes no lineales en los que la función de regresión acepta una aproximación lineal de la forma $f(x \mid \boldsymbol{\theta}) \approx \theta_0 + \sum_{i=1}^{k} \theta_i\, g_i(x)$.

Se concluye el programa describiendo, sin detallar los cálculos, el resultado final de un análisis de regresión complejo con datos reales (predicciones 'on line' en una noche electoral), en el que se han utilizado los métodos bayesianos objetivos estudiados en la asignatura.

7 DISCUSIÓN: UN PROGRAMA POSIBLE

Este trabajo concluye presentando un posible programa para el desarrollo de la asignatura de *Estadística Matemática* en la línea descrita, seguido de algunos comentarios sobre la forma en que podrían organizarse las clases prácticas, y de algunas recomendaciones sobre posibles textos de apoyo.

7.1 Programa Propuesto

El temario que se propone está orientado a una asignatura de *Estadística Matemática* situada en un *primer ciclo* universitario, y supone conocimientos elementales de análisis y de teoría de la probabilidad. En el caso en el que los alumnos no hayan cursado una asignatura de teoría de la probabilidad, será necesario ampliar la primera parte del programa para introducir los conceptos necesarios.

El nivel al que pueda desarrollase el temario propuesto dependerá obviamente del alumnado, resultando presumiblemente más sencillo dotar al curso de un nivel elevado en el caso de contar con alumnos de Matemáticas o de Ciencias Estadísticas.

El programa que sugerimos se estructura en 18 temas, lo que en una asignatura anual con 6 créditos teóricos, como la prevista en varios de los planes de estudio vigentes, permite dedicar algo más de 3 horas lectivas a cada tema.

1. *Fundamentos*
   1. Introducción a Estadística Matemática
   2. Medida de Probabilidad
   3. Introducción a la Teoría de Decisión
   4. Medidas de Información y de Discrepancia
2. *Modelos Probabilísticos*
   5. Intercambiabilidad e Independencia Condicional
   6. Función de Verosimilitud
   7. Suficiencia y Familia Exponencial
3. *Inferencia: Métodos Bayesianos Objetivos*
   8. El Paradigma Bayesiano
   9. Distribuciones de Referencia
   10. Estimación
   11. Contraste de Hipótesis
4. *Calibración: Métodos Frecuentistas*
   12. Comportamiento Medio de un Procedimiento Estadístico
   13. Error Medio de los Estimadores Puntuales
   14. Recubrimiento Esperado de las Regiones de Confianza
   15. Probabilidades de Error en los Contrastes de Hipótesis
5. *Aplicaciones*
   16. Comparación de Proporciones
   17. Inferencia con Medias Normales
   18. Regresión

7.2 Prácticas

La posibilidad actual, presente en casi todas las universidades, de dar las clases prácticas en aulas de informática dotadas de ordenadores razonablemente potentes, permite orientar las clases prácticas a la verificación *experimental*, por simulación, de los conceptos y resultados desarrollados en el curso.

Son muchos los lenguajes de programación que pueden ser utilizados con este propósito, pero nuestra preferencia se sitúa en el uso de *Mathematica*. La experiencia acumulada sugiere que el alumno es capaz en unas pocas horas de manejar el programa con relativa soltura, y que pronto lo utilizará sistemáticamente para otras asignaturas.

Los 3 créditos de prácticas generalmente asociados a la asignatura de Estadística Matemática permiten montar 10 prácticas de 3 horas cada una, suficientes para experimentar con todos los conceptos desarrollados, especialmente si materiales adecuados (preparados ex-profeso) son directamente accesibles a los alumnos mediante una red local.

7.3 Bibliografía Básica

Dado el carácter poco convencional del programa propuesto, no es fácil encontrar textos que se adapten bien a su contenido; sin embargo, la mayor parte de los conceptos expuestos están muy bien descritos en DeGroot & Schervish (2002), la tercera edición de un texto excelente, de cuya primera edición existe una versión en castellano.

El libro de Barnett (1999), una buena introducción al análisis comparativo de las distintas escuelas de inferencia, puede ser un interesante complemento. El artículo sobre inferencia bayesiana de la enciclopedia de la UNESCO (Bernardo, 2002), proporciona una introducción elemental a los métodos bayesianos objetivos poco frecuente en los libros de texto. Schervish (1995), un excelente tratado de estadística teórica con un fuerte contenido matemático, puede resultar muy útil como libro de consulta.

Los conceptos básicos de la teoría de la decisión pueden ser estudiados en la extraordinaria introducción de Lindley (1985), de cuya primera edición también existe una versión en castellano. Los libros de DeGroot (1970) y de Bernardo & Smith (1994) contienen versiones formales de los fundamentos de la estadística matemática, que pueden servir de complemento.

Una buena introducción a los métodos bayesianos objetivos es la contenida en Box & Tiao (1973), un texto clásico que sigue vigente. Berger (1985) y Bernardo & Smith (1994) proporcionan contenidos más modernos.

Los conceptos más convencionales de la estadística matemática pueden estudiarse en los numerosos textos publicados con una orientación básicamente frecuentista. Entre los escritos en castellano, destacamos los de Cao *et al.* (2001), Peña (1999) y Vélez-Ibarrola & García-Pérez (1997). Los textos de Berry & Lindgren (1996), Casella & Berger (2002) y Shao (1999) constituyen interesantes alternativas.

AGRADECIMIENTOS

Investigación financiada con los Proyectos GV01-7 de la Generalitat Valenciana y BNF2001-2889 de la DGICYT, Madrid.

BIBLIOGRAFÍA

1. Barnett. V. (1999). *Comparative Statistical Inference* (Tercera edición). Chichester: Wiley.

2. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer.

3. Berry D. A. & Lindgren, B. W. (1996). *Statistics: Theory and Methods*. Duxbury.

4. Berger, J. O. & Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (con discusión).

5. Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (con discusión). Reeditado en *Bayesian Inference* (N. G. Polson & G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229–263.

6. Bernardo, J. M. (2002). Bayesian Statistics. *Encyclopedia of Life Support Systems (EOLSS)*. Paris: UNESCO (en prensa). Accesible on line: `http://www.uv.es/~bernardo/`.

7. Bernardo, J. M. (1999). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 101–130 (con discusión).

8. Bernardo, J. M. & Juárez, M. (2003). Intrinsic Estimation. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford: University Press, (en prensa).

9. Bernardo, J. M. & Ramón, J. M. (1998). An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *The Statistician* **47**, 1–35.

10. Bernardo, J. M. & Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *J. Statist. Planning and Inference* **70** (en prensa).

11. Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

12. Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley. Reeditado en 1992, New York: Wiley.

13. Cao, R, Francisco, M., Naya, S. Presedo, M. A., Vázquez M., Vilar, J. A. & Vilar, J. M. (2001). *Introducción a la Estadística y sus Aplicaciones*. Madrid: Pirámide.

14. Casella, G. & Berger, R. L. (2002). *Statistical Inference* (Segunda edición). New York: Duxbury.

15. DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

16. DeGroot, M. H. & Schervish, M. J. (2002). *Probability and Statistics* (Tercera edición). Addison-Wesley. Versión española de la primera edición: DeGroot, M. H. (1988). *Probabilidad y Estadística*. Mexico: Addison-Wesley Iberoamericana.

17. de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68. Reimpreso en 1980 como 'Foresight; its logical laws, its subjective sources' en *Studies in Subjective Probability* (H. E. Kyburg and H. E Smokler, eds.). New York: Dover, 93–158.

18. de Finetti, B. (1970). *Teoria delle Probabilità*. Turin: Einaudi. Traducido al inglés como *Theory of Probability* (1975), Chichester: Wiley.

19. Hewitt, E. & Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.

20. Huzurbazar, V. S. (1976). *Sufficient Statistics*. New York: Marcel Dekker.

21. Jeffreys, H. (1939). *Theory of Probability*. Oxford: University Press. Tercera edición (1961) reimpresa en 1998, Oxford: University Press.

22. Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. Reimpreso como *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.

23. Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: University Press.

24. Lindley, D. V. (1985). *Making Decisions*. (Segunda edición) Chichester: Wiley. Versión española de la primera edición: Lindley, D. V. (1977) *Introducción a la Teoría de la Decisión*. Barcelona, Vicens-Vives.

25. Lindley, D. V. & Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *Amer. Statist.* **30**, 112–119.

26. Peña, D. (1999). *Estadística: Modelos y Métodos* (Décima edición). Madrid: Alianza Universidad.

27. Press, S. J. (1972). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Segunda edición en 1982, Melbourne, FL: Krieger.

28. Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley. Segunda edición en 1972, New York: Dover.

29. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 y 623–656. Reimpreso en *The Mathematical Theory of Communication* (Shannon, C. E. & Weaver, W., 1949). Urbana, IL.: Univ. Illinois Press.

30. Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer

31. Shao, J. (1999). *Mathematical Statistics*. New York: Springer

32. Vélez-Ibarrola, R. & García-Pérez, A. (1997). *Principios de Inferencia Estadística*. Madrid: UNED

33. Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley. Reimpreso en 1987, Melbourne, FL: Krieger.