CÓMO HACER INFERENCIA ESTADÍSTICA A PARTIR DE DATOS DE BAJA CALIDAD

Inés Couso
Dep. Estadística e I.O. y D.M.
Universidad de Oviedo
e-mail: couso@uniovi.es

Resumen

En problemas de aprendizaje de máquina, si el conjunto de entrenamiento no es conocido de forma precisa, tampoco es posible determinar de forma exacta el error de un modelo. Por consiguiente, la comparación estadística entre diferentes algoritmos de aprendizaje no puede resolverse con los diseños experimentales habituales. Este trabajo estudia las extensiones de la inferencia estadística y las limitaciones de la toma de decisiones a los problemas con datos de baja calidad.

Palabras Clave: Datos de baja calidad, Diseño experimental, Inferencia estadística.

1. INTRODUCCIÓN

La última fase del diseño de un algoritmo de aprendizaje es la validación estadística de sus resultados. Los diseños experimentales más extendidos emplean tests para contrastar la precisión, interpretabilidad, velocidad de ejecución, u otro índice de calidad del algoritmo.

En general, dichos índices son atributos numéricos. Sin embargo, cuando los datos de entrenamiento son intervalares, o borrosos, los índices con los que se mide la calidad de un modelo son, a su vez, intervalos de valores o números borrosos. No tenemos constancia de que, hasta el momento, se haya detallado una metodología de comparación de algoritmos válida para este caso. Por esta razón, en este trabajo se estudiarán los procesos de inferencia estadística necesarios para comparar los resultados de los modelos que se obtienen cuando los datos de aprendizaje son de baja calidad.

Nuestra metodología parte de la base de que se debe utilizar toda la información disponible, pero sin añadir ningún dato de forma artificial. El hecho de añadir información artificial, para poder aprovechar técnicas estadísticas ya existentes, puede hacernos incurrir en conclusiones erróneas. No obstante, a lo largo de este trabajo haremos las siguientes consideraciones:

- Un atributo sigue cierta distribución de probabilidad en una población.
- Dicha distribución es conocida salvo por el valor de un parámetro (esto es, no consideraremos tests no paramétricos).
- Se dispone de n intervalos que representan la información (imprecisa) disponible acerca del valor del atributo en esas n mediciones (tal y como ocurriría, por ejemplo, si se hubiese realizado un diseño experimental basado en n-cv).

Para aclarar los conceptos introducidos en el trabajo utilizaremos ejemplos que involucran cálculos sencillos. Aunque no son ejemplos reales, los utilizamos deliberadamente, para que el lector no desvíe su atención hacia el problema de la complejidad del cálculo numérico. Por último, indicamos que en este estudio nos restringiremos a aquel caso en que nuestra información muestral viene dada por una colección de intervalos. La extensión al caso difuso es posible, pero por razones de espacio no se ha incluido aquí.

La estructura de este trabajo es como sigue: en la sección 2, explicaremos los dos tratamientos teóricos existentes en la literatura para los datos intervalares, y mostraremos cuál de estos tratamientos se corresponde con nuestra noción de datos de baja calidad. En la sección 3, explicamos cómo debe extenderse la noción clásica de test a nuestro contexto. En la sección 4, trataremos el problema de la comparación de la potencia de distintos tests. El trabajo termina con las conclusiones y el trabajo futuro en la sección 5.

2. DOS INTERPRETACIONES DE LOS DATOS INTERVALARES

A lo largo de esta sección, partiremos de la información que nos proporciona la observación (incompleta) en dos muestras apareadas. Así, nuestras entradas serán una colección de n parejas de intervalos, tal como se muestra en el cuadro 1:

Cuadro 1: Información muestral

muestra 1	$[l_1, u_1]$	...	$[l_n, u_n]$
muestra 2	$[l'_1, u'_1]$	...	$[l'_n, u'_n]$

Hay dos grandes grupos de técnicas estadísticas aplicables a datos intervalares: aquellas en las que se asume que el intervalo representa información imprecisa acerca de un valor real conocido parcialmente y aquellas en las que se asume que las salidas del experimento aleatorio son realmente intervalos. Ambas interpretaciones son válidas, pero cada una es aplicable en un contexto distinto. Con la ayuda de dos ejemplos sencillos, mostraremos la utilidad de cada una. En ambos ejemplos, partiremos de la misma información muestral, pero el tratamiento de los datos deberá ser diferente.

Ejemplo 2.1 *Se pretende comparar las temperaturas ambientales de las localidades de Huelva y Adeje (sur de Tenerife) durante los meses de mayo de los últimos 25 años. Con ese fin, se han tomado los datos del histórico para 36 días elegidos al azar. En el cuadro 2 se muestran los intervalos [Tmin, Tmax] en ambas localidades.*

Cuadro 2: Temperaturas máximas y mínimas.

Huelva	[15,25]	[15,25]	[16,25]	[18,27]
Adeje	[19,26]	[20,25]	[21,26]	[21,26]
nº días	10	5	12	9

Denotemos por μ_{mH} la temperatura mínima media (esperanza de las mínimas) y por μ_{MH} la temperatura máxima media en la ciudad de Huelva. De forma análoga, denotemos por μ_{mA} y por μ_{MA}, respectivamente, a las cantidades análogas en la localidad de Adeje. Nuestro objetivo es contrastar si los intervalos $[\mu_{mH}, \mu_{MH}]$ y $[\mu_{mA}, \mu_{MA}]$ son similares. Para hacer un contraste de este tipo, un procedimiento válido[1] *sería el siguiente: en primer lugar, elegimos una métrica adecuada para cuantificar distancias entre intervalos. Por ejemplo, podemos elegir la distancia:*

$$d([a,b],[c,d]) = \sqrt{\frac{(c-a)^2 + (d-b)^2}{2}}.$$

que pertenece a la familia de distancias propuesta por Bertoluzza et al. en [1]. A partir de los datos anteriores, las distancias entre intervalos son:

Cuadro 3: Distancias entre intervalos

distancia	$\sqrt{5}$	$\sqrt{8{,}5}$	$\sqrt{12{,}5}$	$\sqrt{13}$
nº días	9	10	5	12

A continuación, se contrastaría si el valor esperado de las distancias es 0 o no. A partir de los datos recogidos, rechazaríamos la hipótesis nula (omitiremos los cálculos necesarios para llegar a esa conclusión). En efecto, aunque las temperaturas máximas en mayo son parecidas en ambas localidades, las temperaturas mínimas no lo son, ya que el clima en el sur de Tenerife es más suave.

Ejemplo 2.2 *Ahora pretendemos contrastar si las temperaturas ambientales de ambas localidades a las 16:00 h (horas locales) son similares o no. Para los mismos 36 días del ejemplo anterior, se han recogido las parejas de temperaturas a esa hora en ambas localidades. Los datos se presentan en el cuadro 4:*

Cuadro 4: Temperaturas a las 16:00h.

Huelva	23	24	25	25	26
Adeje	24	25	24	25	25
nº días	4	13	11	4	4

La muestra de las diferencias entre las temperaturas de ambas localidades en los 36 días señalados se muestra en el cuadro 5.

Cuadro 5: Diferencias a las 16:00h

diferencias	-1	0	1
nº días	17	4	15

No podemos descartar la hipótesis (omitimos los cálculos, por falta de espacio) de que esta muestra proviene de una variable aleatoria con media nula. Por

[1] En la literatura reciente podemos encontrar algunos trabajos dedicados a la determinación de tests para problemas en que los datos muestrales son intervalos interpretados de la forma indicada aquí. En todos ellos está implícita la idea de contrastar simultáneamente la igualdad entre las esperanzas de los extremos de los respectivos intervalos ($H_0 : \mu_{mH} = \mu_{mA}$ y $\mu_{MH} = \mu_{MA}$), o, de forma equivalente, contrastar la igualdad entre las esperanzas de los puntos medios y de las amplitudes ($H_0 : \frac{\mu_{mH}+\mu_{MH}}{2} = \frac{\mu_{mA}+\mu_{MA}}{2}$ y $\frac{\mu_{MH}-\mu_{mH}}{2} = \frac{\mu_{MA}-\mu_{mA}}{2}$.) Está fuera de los propósitos de nuestro trabajo hacer un análisis exhaustivo de estos procedimientos.

tanto, si denotamos respectivamente por μ_1 y μ_2 a las temperaturas medias en los días de mayo a las 16:00 h en ambas localidades, no rechazaremos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_0 : \mu_1 \neq \mu_2$.

Supongamos ahora, que una persona debe contrastar ambas hipótesis, pero que no dispone de los datos del cuadro 5, sino sólo de las temperaturas máximas y mínimas diarias recogidas en el cuadro 2. Así que su única información acerca de la temperatura a las 16:00 h en un día concreto es que se encuentra entre los valores máximo y mínimo registrados. (Supongamos que no tiene ninguna información adicional acerca de la relación entre la temperatura a las 16:00 h y las temperaturas máxima y mínima del día). Con una información tan imprecisa, la persona no puede contrastar las hipótesis anteriores. En efecto, para la muestra real de las temperaturas (que él desconoce), no rechazaríamos la hipótesis nula. Sin embargo, la muestra siguiente también es compatible con los datos del cuadro 2:

Cuadro 6: Muestra compatible alternativa

Huelva	15	15	16	18
Adeje	26	25	26	26
nº días	10	5	12	9

Para esta muestra de datos sí rechazaríamos la hipótesis H_0.

En nuestro contexto, cada intervalo refleja, al igual que en este último ejemplo, nuestra información (incompleta o imprecisa) acerca de un número. El hecho de que las amplitudes y/o los puntos medios de los intervalos de ambas muestras sean, en general, diferentes, no implica que las diferencias entre los valores reales sean significativas. De forma análoga, si la imprecisión es grande, el hecho de que los intervalos de ambas muestras sean parecidos no implica que los datos reales provengan de poblaciones similares. Para tomar decisiones a partir de datos de baja calidad, se considera un test determinado y se calcula su salida (1=rechazo / 0=no rechazo) para cada colección de n parejas compatible con la información disponible en el cuadro 2. Nos encontraremos en alguno de los tres casos siguientes:

- Para todas ellas, el test indica "rechazar". (Rechazaremos H_0).
- Para todas ellas, el test indica "no rechazar". (No rechazamos la hipótesis nula).
- Para algunas muestras, el test indica "rechazar" y para otras, "no rechazar". En ese caso, la imprecisión de los datos nos impide tomar una decisión clara, porque no se dispone de información suficiente.

En la sección 3, introduciremos algunas definiciones formales en relación con este problema.

2.1. Aplicación a la verificación en aprendizaje de máquina

Cabe preguntarse cuál de los dos enfoques es el más adecuado para la comparación de algoritmos de aprendizaje a partir de datos de baja calidad. Como hemos mencionado, en la validación de un algoritmo de aprendizaje se dispone de una muestra de n intervalos, cada uno de los cuales restringe los valores posibles del atributo que se desee contrastar. Con frecuencia, se busca un test que permita rechazar que los valores medios de los atributos obtenidos para dos algoritmos sean iguales, y por esta razón la segunda técnica es la más adecuada; si se emplease la primera, se estaría contrastando si las medias de los márgenes superiores e inferiores de los valores del atributo coinciden. En general, esta última información no es necesaria para la validación.

3. LIMITACIONES EN LA TOMA DE DECISIONES

En esta sección, vamos a mostrar una definición formal que generaliza el concepto de *test* al caso que nos ocupa. Antes debemos recordar algunas nociones de la estadística clásica. Consideremos una muestra aleatoria simple $(\mathbf{X}, \mathbf{Y}) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ tomada a partir de un vector aleatorio bidimensional (X, Y) (representa n mediciones tomadas al azar de la pareja de características representada por (X, Y)). (En el ejemplo 2.2, X representaría la temperatura en Huelva e Y, la temperatura en Tenerife y $((X_1, Y_1), \ldots, (X_n, Y_n))$ denotaría la colección de $n = 36$ parejas de mediciones realizadas.) Denotemos por θ_X (respectivamente por θ_Y) a cierto parámetro (esperanza, mediana, etc.) asociado a la distribución de X (respectivamente, a la distribución de Y). Denotemos ahora por $\varphi : (\mathbb{R}^2)^n \to \{0, 1\}$ a una función de test para contrastar

$$H_0 : \theta_X = \theta_Y \quad \text{frente a} \quad H_1 : \theta_X \neq \theta_Y.$$

La función φ asocia, a cada posible realización muestral, $(\mathbf{x}, \mathbf{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathbb{R}^2)^n$, el valor 1

(rechazo) o 0 (no rechazo). Denotemos por R a la región crítica asociada a φ,

$$R = \{(\mathbf{x}, \mathbf{y}) \in (I\!R^2)^n \,:\, \varphi(\mathbf{x}, \mathbf{y}) = 1\}.$$

Llamamos *tamaño* del test φ a la cantidad $\sup_{\theta\in\Theta_0} E_\theta(\varphi(\mathbf{X})) = \sup_{\theta\in\Theta_0} P_\theta(\text{Rechazar} H_0)$. Supongamos ahora que, para cada realización del experimento se obtienen una pareja de cotas de los valores de ambas variables. Consideremos, por tanto, una tupla de n vectores, $((L_1^X, U_1^X, L_1^Y, U_1^Y), \ldots, (L_n^X, U_n^X, L_n^Y, U_n^Y))$ donde L_i^X (respectivamente L_i^Y) denota la cota inferior para el valor correspondiente a la i-ésima medición y U_i^X (respectivamente U_i^Y) denota la cota superior. Dicha tupla representa nuestra información (incompleta) acerca de la m.a.s. $((X_1, Y_1), \ldots, (X_n, Y_n))$. De esta forma, para una realización muestral concreta, nuestra información vendrá dada por una colección ordenada de $4n$ valores de la forma $\gamma = ((l_1^X, u_1^X, l_1^Y, u_1^Y), \ldots, (l_n^X, u_n^X, l_n^Y, u_n^Y))$ La salida del test φ será la siguiente:

$$\varphi(\gamma) = \{\varphi(x_1, \ldots, x_n) \,:\, l_i \le x_i \le u_i,\ \forall\, i = 1, \ldots, n\}.$$

Siempre que el parámetro θ sea un operador lineal, tal como ocurre con la esperanza, el contraste anterior se puede expresar alternativamente de la forma $H_0 : \theta_{X-Y} = 0$. Así, si denotamos por Z a la variable diferencia, $Z = X - Y$, el contraste anterior representa un caso particular de $H_0 : \theta_Z = \theta_0$ frente a $H_1 : \theta_Z \in \Theta_1$, donde $\Theta_1 \not\ni \theta_0$. A partir de ahora nos referiremos a este contraste más general y utilizaremos la notación θ en lugar de θ_Z, lo cual no dará lugar a confusión. Así, en lo que resta de trabajo, supondremos que nuestra entrada es un producto cartesiano $\gamma = [l_1, u_1] \times \ldots \times [l_n, u_n]$, que representa nuestra información imprecisa acerca de una tupla de n valores, $\mathbf{z} = (z_1, \ldots, z_n)$, tomados a partir de la variable $Z = X - Y$.

Dado un test, φ, para el contraste de la hipótesis $H_0 : \theta_Z = \theta_0$ frente a $H_1 : \theta_Z \in \Theta_1$, sus posibles salidas serán las siguientes:

$$\varphi(\gamma) = \begin{cases} \{1\} \text{ (rechazo)} & \text{si } \gamma \subseteq R \\ \{0\} \text{ (no rechazo)} & \text{si } \gamma \subseteq I\!R^n \setminus R \\ \{0, 1\} \text{ (indecisión)} & \text{en otro caso,} \end{cases}$$

donde R denota la región crítica asociada al test φ, $R = \{\mathbf{x} \,:\, \varphi(\mathbf{x}) = 1\}$.

Observemos que esta forma de proceder es conservativa, en el sentido de que si partimos de un test de tamaño α, la probabilidad de rechazo ($\varphi(\gamma) = \{1\}$) está acotada por dicho valor α, aun cuando los datos son imprecisos.

4. ELECCIÓN DE TESTS

Dado un test φ, la potencia asociada a un valor $\theta \in \Theta_1$, $\text{pot}_\varphi(\theta) = P_\theta(\varphi(\mathbf{X}) = 1)$, representa la probabilidad de rechazar la hipótesis nula, bajo la suposición de que el valor del parámetro es θ. Cuando las observaciones muestrales son totalmente precisas, un criterio extendido en la práctica consiste en elegir, entre todos los tests de un tamaño prefijado, α, aquel test φ^*(si es que tal existe) que sea uniformemente más potente, es decir, que satisfaga las desigualdades:

$$\text{pot}_{\varphi^*}(\theta) \ge \text{pot}_\varphi(\theta),\ \forall\, \theta \in \Theta_1, \forall\, \varphi.$$

Pero, cuando la información muestral es de tipo intervalar, el test uniformemente más potente no es, necesariamente, el que mejores propiedades tiene en la práctica. En una primera aproximación al problema de encontrar un criterio para comparar tests en este nuevo contexto, podemos centrar nuestra atención en las cantidades: $\underline{\text{pot}}(\theta) = P_\theta(\varphi(\Gamma) = \{1\})$ (probabilidad, bajo el supuesto de que el verdadero valor del parámetro es θ, de que el test indique "rechazar") y $\overline{\text{pot}}(\theta) = P_\theta(\varphi(\Gamma) \ni 1)$ (probabilidad de rechazar o no concluir, bajo el mismo supuesto). Intuitivamente, cuanto más altas sean estas cantidades, más adecuado será el test φ. Pero surgen varias dificultades, que enumeramos a continuación:

1. $\underline{\text{pot}}(\theta)$ y $\overline{\text{pot}}(\theta)$ no son, realmente dos valores bien definidos en situaciones no triviales. En efecto, las probabilidades de las respectivas posibles salidas del test (rechazar, no rechazar, indecisión), dependen, en última instancia, de las distribuciones de probabilidad de lo que nosotros observamos (U=temperatura máxima diaria" y L="temperatura mínima diaria") y, por tanto, no quedan determinadas aunque conozcamos completamente la distribución de la variable Z (temperatura a las 16:00h). El problema anterior tiene una solución parcial. Aunque no disponemos de la información suficiente acerca de la relación entre nuestra información (observaciones de L y de U) y los valores reales de la variable X, como para determinar la distribución del vector (L, U) a partir de un determinado θ, sí tenemos, en ocasiones, cierta información (incompleta) acerca de esta relación. Así, en general, podemos calcular cotas inferiores y superiores no triviales (es decir, distintas de 1 y 0, respectivamente) para $\overline{\text{pot}}(\theta)$ y para $\underline{\text{pot}}(\theta)$.

   A continuación mostramos un ejemplo sencillo para aclarar esta idea:

Ejemplo 4.1 *Consideremos una v.a. Z con distribución $N(\mu,1)$[2], donde μ es un valor desconocido. Supongamos que no podemos observar valores de Z de forma precisa, sino que únicamente podemos observar 25 valores tomados al azar a partir de una variable Y, de la cual sabemos que cumple la restricción $|Y-Z| \leq 0{,}1$. Queremos contrastar la hipótesis $H_0 : \mu = 0$ frente a $H_1 : \mu < 0$. Bajo la hipótesis nula, la variable aleatoria $\overline{Z}$ (media muestral) sigue distribución N(0,0.2). Por lo tanto, el test*

$$\varphi(\mathbf{z}) = \begin{cases} 1 \text{ (rechazo)} & \text{si } \overline{z} < -0{,}33 \\ 0 \text{ (no rechazo)} & \text{si } \overline{z} \geq -0{,}33, \end{cases}$$

es un test de tamaño $\alpha = 0{,}05$ para este contraste. Para una realización muestral, sabemos que cada z_i $(i = 1, \ldots, 25)$ pertenece al intervalo $[y_i - 0{,}1, y_i + 0{,}1]$. Así que, para una secuencia de observaciones $\mathbf{y} = (y_1, \ldots, y_{25})$ la salida del test será la siguiente:

$$\varphi(\gamma) = \begin{cases} \{1\} \text{ (rechazo)} & \text{si } \overline{y} < -0{,}43 \\ \{0\} \text{ (no rechazo)} & \text{si } \overline{y} \geq -0{,}23 \\ \{0,1\} \text{ (indecisión)} & \text{en otro caso.} \end{cases}$$

Para el valor $\mu = -1$ la potencia de este test es:

$$\text{pot}(-1) = P_{\mu=-1}(\overline{Z} < -0{,}33) =$$

$$P_{\mu=-1}(5(\overline{Z}+1) < 3{,}3) = \Phi_{N(0,1)}(3{,}35) = 0{,}9996.$$

Pero este valor nos informa de la probabilidad de rechazar H_0 (bajo el supuesto de que $Z \equiv N(-1,1)$ a partir de una secuencia de 25 observaciones aleatorias totalmente precisas. No nos proporciona la probabilidad de rechazar a partir de una secuencia de 25 observaciones imprecisas.

Observemos que las probabilidades

$$P(\{\gamma : \varphi(\gamma) = \{1\}\}) = P(\overline{Y} < -0{,}43) \text{ y}$$

$$P(\{\gamma : \varphi(\gamma) = \{1\} \text{ ó } \varphi(\gamma) = \{0,1\}\}) =$$

$$P(\overline{Y} \leq -0{,}23)$$

no quedan completamente especificadas a partir del supuesto de que $Z \equiv N(-1,1)$. Pero teniendo en cuenta las desigualdades:

$$\overline{Z} - 0{,}1 \leq \overline{Y} \leq \overline{Z} + 0{,}1,$$

podemos afirmar que:

$$P(\overline{Z}+0{,}1 < y) \leq P(\overline{Y} < y) < P(\overline{Z}-0{,}1 < y), \ \forall\, y.$$

Así, el valor: $P_{\mu=-1}(\overline{Z} + 0{,}1 < -0{,}43) = \Phi_{N(0,1)}(2{,}35) = 0{,}9906$ es una cota inferior para $P(\{\gamma : \varphi(\gamma) = \{1\}\})$, mientras que el valor $P_{\mu=-1}(\overline{Z} + 0{,}1 < -0{,}23) = \Phi_{N(0,1)}(4{,}35) \approx 1$ es una cota superior para $P(\{\gamma : \varphi(\gamma) = \{1\}$ o $\varphi(\gamma) = \{0,1\}\})$. Por otra parte, la cantidad $\text{pot}(-1) = 0{,}9995$ es, simultáneamente, una cota superior para la primera y una cota inferior para la segunda.

2. Sin embargo, aunque esas cotas estén bien definidas, su cálculo es un problema de difícil resolución, tal como mostramos en el siguiente ejemplo.

Ejemplo 4.2 *Consideremos ahora una v.a. Z con distribución $N(\mu,\sigma)$, donde μ y σ son valores desconocidos. Supongamos que no podemos observar valores de Z de forma precisa, sino que únicamente podemos observar 25 valores tomados al azar a partir de una variable Y, de la cual sabemos que cumple la restricción $|Y-Z| \leq 0{,}1$. Queremos contrastar la hipótesis $H_0 : \mu = 0$ frente a $H_1 : \mu < 0$. Bajo la hipótesis nula, la variable aleatoria $5\overline{Z}/S_Z$, donde S_Z denota la cuasi-varianza muestral, sigue distribución t_{24}. Por lo tanto, el test*

$$\varphi(\mathbf{z}) = \begin{cases} 1 \text{ (rechazo)} & \text{si } \frac{5\overline{z}}{s_z} < -1{,}71 \\ 0 \text{ (no rechazo)} & \text{si } \frac{5\overline{z}}{s_z} \geq -1{,}71, \end{cases}$$

es un test de tamaño $\alpha = 0{,}05$ para este contraste. Para una realización muestral, sabemos que cada z_i $(i = 1, \ldots, 25)$ pertenece al intervalo $[y_i - 0{,}1, y_i + 0{,}1]$. Así que, para una secuencia de observaciones $\mathbf{y} = (y_1, \ldots, y_{25})$ la salida del test será la siguiente:

$$\varphi(\gamma) = \begin{cases} \{1\} & \text{si } \sup_{\mathbf{x}\,:\,x_i \in [y_i-0{,}1, y_i+0{,}1] \forall i} \frac{5\overline{x}}{s_x} < -1{,}71 \\ \{0\} & \text{si } \inf_{\mathbf{x}\,:\,x_i \in [y_i-0{,}1, y_i+0{,}1] \forall i} \frac{5\overline{x}}{s_x} > -1{,}71 \\ \{0,1\} & \text{en otro caso.} \end{cases}$$

Dado un vector $\mathbf{y} = (y_1, \ldots, y_n)$ la determinación de las cantidades $sup_{\mathbf{x}\,:\,x_i \in [y_i-0{,}1, y_i+0{,}1] \forall i} s_x$ e

[2]En este ejemplo no pretendemos ilustrar una situación real. No son habituales los problemas en los que se desconoce la media de la población, pero se conoce la varianza. Por otra parte, tampoco mostramos las dificultades que presenta el hecho de determinar si los datos muestrales provienen o no de una población normal, cuando son observados de forma imprecisa.

$\inf_{\mathbf{x}\,:\,x_i\in[y_i-0{,}1,y_i+0{,}1]\forall\, i}\, s_x$ *no es un problema trivial y ha sido objeto de diversos estudios recientes (ver, por ejemplo, [6]). Por tanto, en este ejemplo, la determinación de la salida del test para cada posible muestra imprecisa no es inmediata. Así, aunque, en principio, existen cotas no triviales para* $\underline{\mathrm{pot}}(\theta)$ *y* $\overline{\mathrm{pot}}(\theta)$ *(Denotemos por* $\underline{\mathrm{pot}_*}_{\varphi}(\theta)$ *a una cota inferior no trivial de* $\underline{\mathrm{pot}}_{\varphi}(\theta)$ *y por* $\overline{\mathrm{pot}^*}_{\varphi}(\theta)$ *a una cota superior de* $\overline{\mathrm{pot}}_{\varphi}(\theta)$*, respectivamente), su determinación requiere la utilización de ciertos algoritmos de optimización.*

3. Para comparar dos tests diferentes, φ_1 y φ_2, es necesario comparar los intervalos de cotas asociados, $[\underline{\mathrm{pot}_*}_{\varphi_1}(\theta), \overline{\mathrm{pot}^*}_{\varphi_1}(\theta)]$ y $[\underline{\mathrm{pot}_*}_{\varphi_2}(\theta), \overline{\mathrm{pot}^*}_{\varphi_2}(\theta)]$. Si $\underline{\mathrm{pot}_*}_{\varphi_1}(\theta) \geq \overline{\mathrm{pot}^*}_{\varphi_2}(\theta)$, preferiremos el test φ_1 al test φ_2. Análogamente, $\underline{\mathrm{pot}_*}_{\varphi_2}(\theta) \geq \overline{\mathrm{pot}^*}_{\varphi_1}(\theta)$, preferiremos el segundo test. Pero en general, los intervalos $(\underline{\mathrm{pot}_*}_{\varphi_1}(\theta), \overline{\mathrm{pot}^*}\varphi_1(\theta))$ y $(\underline{\mathrm{pot}_*}_{\varphi_2}(\theta), \overline{\mathrm{pot}^*}\varphi_2(\theta))$ tendrán intersección no vacía, de forma que habrá que elegir un criterio adecuado para compararlos.

Las dificultades señaladas en los tres puntos anteriores serán objeto de nuestro estudio en un futuro próximo.

5. CONCLUSIONES

En este trabajo nos hemos ceñido al caso en que podemos modelar nuestra información por medio de intervalos. Este tipo de información puede encuadrarse en el marco más general de la teoría de la posibilidad. En efecto, la distribución de posibilidad 0-1-valuada:

$$\pi(c) = \begin{cases} 1 & \text{si } c \in [l,u] \\ 0 & \text{en otro caso,} \end{cases}$$

indica que el grado de seguridad (necesidad) de que el valor desconocido z pertenece al valor $[l,u]$ es 1, ya que la medida de necesidad asociada a ella satisface la igualdad: $N([l,u]) = 1 - \Pi([l,u]^c) = 1 - \sup_{c\in[l,u]^c} \pi(c) = 1$. Así, la nociones de test, p-valor e intervalo de confianza han sido extendidas recientemente al caso en que los datos muestrales (entradas) son subconjuntos difusos de $\mathbb{R}^n$ que indican distribuciones de posibilidad más generales. Tal como se demuestra en [2], dichos grados de posibilidad (o valores de la función de pertenencia) quedan completamente especificados a partir de una familia anidada de intervalos de confianza que describe la información disponible acerca del valor z. En recientes trabajos, se han estudiado las extensiones de los conceptos de test, de p-valor y región de confianza a este contexto más general ([3, 4, 5]). En concreto, el p-valor extendido se define como un conjunto difuso, que se interpreta como una medida de posibilidad. En [3] hemos reducido ese conjunto difuso a un intervalo, dotándole de una interpretación coherente dentro de la teoría de las probabilidades imprecisas [7]. En un futuro próximo nos proponemos extender el concepto de potencia a este contexto más general. Presumiblemente, dicha extensión dará lugar a un número difuso, que podrá interpretarse como una medida de posibilidad. Al igual que en el caso del p-valor, intentaremos reducir este difuso a un intervalo, de forma que el problema de encontrar un test adecuado no se complique aun más para ese caso.

Agradecimientos

Este trabajo está financiado por el Ministerio de Educación y Ciencia, proyectos TIN2007-67418-C03-03 y MTM2007-61193, y por fondos FEDER FEDER-MEC.

Referencias

[1] C. Bertoluzza, N. Corral, A. Salas, On a new class of distances between fuzzy numbers, Mathware Soft Comput. 2 (1995) 71- 84.

[2] I. Couso, S. Montes y P. Gil, The necessity of the strong $\alpha-$cuts of a fuzzy set, IJUFKBS 9 (2001) 249–262.

[3] I. Couso, L. Sánchez, Defuzzification of fuzzy p-values En: Soft Methods for Handling Variability and Imprecision, D. Dubois et al. (Eds.) Springer, 2008.

[4] I. Couso, L. Sánchez, Inner and outer regions induced by low quality data, Eurofuse Workshop'09 -Preference modelling and decision analysis, Navarra, 2009.

[5] T. Denoeux, M.H. Masson, P.A. Hébert, Nonparametric rank-based statistics and significance tests for fuzzy data, FSS 153 (2005) 1–28.

[6] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf and L. Ginzburg, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, T.R. SAND2007-0939, 2007.

[7] P. Walley, Statistical Reasoning with Imprecise Probabilities. 1991. Chapman and Hall, London.