



Universitat d'Alacant
Universidad de Alicante

Muerte en UCIP estimada con el índice "PRISM": comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística y una red neuronal artificial. Una propuesta bayesiana

Vicent Modesto i Alapont

Tesis

Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Muerte en UCIP estimada con el índice "PRISM": comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística y una red neuronal artificial. Una propuesta bayesiana.



Universitat d'Alacant
Universidad de Alicante

*Projecto que, para optar al grado de Doctor,
presenta D. VICENT MODESTO i ALAPONT*

DIRECTORES: Dr. Jaime Latour Pérez, Dr Andreu Nolasco Bonmatí¹ y Dr. Antonio J Serrano López²

¹DEPARTAMENT D'INFERMERIA COMUNITÀRIA,
MEDICINA PREVENTIVA, SALUT PÚBLICA
I HISTÒRIA DE LA CIÈNCIA.
UNIVERSITAT D'ALACANT

Y

²DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA.
UNIVERSITAT DE VALÈNCIA



Universitat d'Alacant
Universidad de Alicante

Dedicada a la memoria de mi abuelo, el Dr. D. José Alapont Llácer. La lectura de su tesis "*Topografía médica de Villanueva de Castellón (Valencia), Año 1948*" ha constituido el estímulo que necesitaba para completar la mía.

Dedicada también a mis padres, Vicente y Josefina, con cariño, por habérmelo dado todo. A mis maestros Jesús Ribera, Lambert Climent y Pascual Malo, con admiración, por haber constituido mis referentes en la vida y en la profesión. Y a mis amigos Emilio, Paco y Silvia, con gratitud, porque me ayudáis cada día a ser mejor persona.

Y, sobretodo, dedicada a Begoña, Mireia y Marta, con todo mi amor, en agradecimiento a todo el tiempo que os he robado para terminarla.



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

"Los hechos previstos, no pueden ser sino probables. Por sólidamente fundada que pueda parecer una predicción, no estamos jamás absolutamente seguros de que la experiencia futura no la desmentirá. (...) Todas las veces que razona por inducción (el científico) hace uso, más o menos conscientemente, del cálculo de probabilidades"¹.

(Jules Henri Poincaré, 1902)

"Medicine is a science of uncertainty and an art of probability".
(Sir William Osler, 1904)

"Observamos y examinamos lo que descubrimos, pero no podemos predecir sin riesgo a equivocarnos. Las posibilidades más razonables a menudo resultan ser falsas"².

(Richard P. Feynman, 1965)

"Lo difícil es predecir, sobretodo el futuro"
(Niels Henrik David Bohr, 1960)

"(...) los medicos deben contentarse en concluir no en certezas, sino en probabilidades estadísticas. El médico moderno, en cierto modo, tiene alguna razón para sentir certeza dentro de los límites estadísticos, pero nunca absoluta seguridad. La certeza absoluta persiste sólo para algunos teólogos y para algunos médicos que aún piensan de igual manera"³

(David Spodick, 1975)



Universitat d'Alacant
Universidad de Alicante



I. - Índice.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

I. - Índice	7
II.- Justificación: introducción y estado actual del tema	15
1.- La predicción en ciencia y el llamado "problema de la inducción".....	17
2.- La teoría de la probabilidad y la estadística como vías para solventar el problema.....	18
3.- La evasión frecuentista: comportamiento inductivo basado en inferencia deductiva.....	20
3.1.- Uso de la evasión frecuentista en la medicina clínica.....	28
4.- La evasión bayesiana: aprendizaje racional desde los datos de la experiencia.....	28
4.1.- La estadística bayesiana como estudio del manejo racional de la incertidumbre.....	29
4.2.- La probabilidad como medida (subjetiva) de la incertidumbre o grado de creencia racional.....	31
4.3.- La evasión bayesiana del problema de la inducción.....	35
4.3.1.- El teorema de Bayes para eventos simples.....	41
4.3.2.- La predicción bayesiana.....	43
4.4.- La distribución "a priori" y la ilusión de objetividad en ciencia.....	45
4.4.1.- Análisis bayesiano de referencia.....	51
4.5.- Inferencia bayesiana: la teoría matemática de la decisión.....	54
5.- Punto de encuentro de ambas evasiones: el concepto de intercambiabilidad y el teorema de representación.....	56
6.- Los índices de predicción clínica como sustento de la toma de decisiones en medicina.....	63
7.- Uso de la evasión bayesiana en medicina clínica: justificación de esta tesis doctoral.....	67
III.- Objetivos	69
IV.- Hipótesis	73
V.- Pacientes, material y métodos	77
1.- La puntuación en el índice "PRISM".....	79
2.- Diseño del estudio.....	80
3.- 1ª Fase: desarrollo de los tests.....	81

3.1- Modelo predictivo de Regresión Logística.....	81
3.1.1.- El diseño.....	81
3.1.2.- Muestreo y tamaño muestral.....	81
3.1.3.- Ajuste del modelo predictivo de regresión logística.....	82
3.1.4.- Evaluación de la exactitud diagnóstica del modelo.....	82
3.1.4.1.- Capacidad discriminante.....	83
3.1.4.2.- Calibración del modelo.....	86
3.2- Red Neuronal Artificial.....	89
3.2.1.- El diseño.....	89
3.2.2.- Muestreo y tamaño muestral.....	89
3.2.3.- Desarrollo de la red neuronal artificial.....	89
3.2.3.1.- Software empleado.....	89
3.2.3.2.- Selección del mejor modelo de RN.....	91
3.2.4.- Evaluación de la exactitud diagnóstica del modelo.....	92
3.2.4.1.- Capacidad discriminante.....	92
3.2.4.2.- Calibración del modelo.....	93
4.- 2ª Fase: validación de ambos tests.....	94
4.1- Diseño.....	94
4.2.- Muestreo.....	94
4.3.- Validación del modelo de regresión logística.....	94
4.4.- Validación del modelo de red neuronal artificial.....	95
4.5.- Evaluación de la exactitud diagnóstica de ambos modelos.....	95
4.5.1.- Evaluación de la capacidad discriminante.....	96
4.5.2.- Evaluación de la calibración.....	96
5.- 3ª Fase: comparación de la exactitud diagnóstica de ambos tests.....	96
5.1.- Comparación de la capacidad discriminante.....	96
5.1.1.- Evaluación clásica.....	97
5.1.2.- Propuesta de evaluación bayesiana.....	97
a) Utilizando la aproximación Normal a la distribución Beta.....	97
b) Utiliando la simulación con métodos de Monte Carlo.....	98
c) Utiliando el método gráfico discreto de la parrilla de Berry.....	99
5.2.- Comparación mediante la incorporación al proceso diagnóstico: análisis de decisión.....	100
VI.- Resultados.....	103

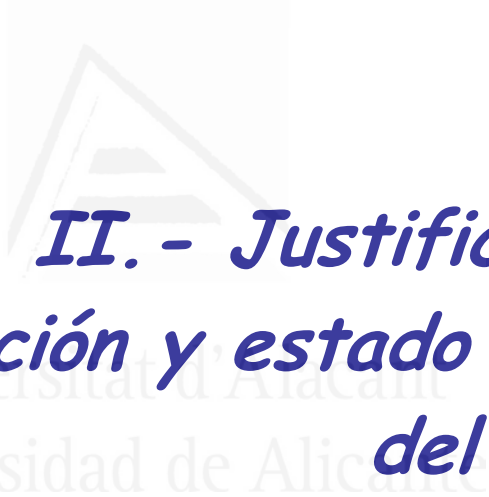
1.- 1ª Fase: desarrollo de los tests.....	105
1.1- Modelo predictivo de Regresión Logística.....	105
1.1.1.- Evaluación clásica de la exactitud diagnóstica.....	106
1.1.1.1.- Capacidad discriminante.....	106
a) Estudio de la curva ROC.....	107
b) Índices de exactitud diagnóstica.....	107
1.1.1.2.- Calibración del modelo.....	108
1.1.2.- Evaluación bayesiana de la exactitud diagnóstica	108
1.1.2.1.- Capacidad discriminante.....	108
a) Estudio bayesiano de la curva ROC.....	108
b) Índices bayesianos de exactitud diagnóstica.....	110
1.1.2.2.- Calibración del modelo.....	117
a) Usando como estimador puntual por estrato el estimador bayesiano convencional.....	118
b) Usando como estimador puntual por estrato el estimador intrínseco.....	119
1.2- Modelo de Red Neuronal Artificial.....	120
1.2.1.- Evaluación clásica de la exactitud diagnóstica.....	124
1.2.1.1.- Capacidad discriminante.....	124
a) Estudio de la curva ROC.....	124
b) Índices de exactitud diagnóstica.....	125
1.2.1.2.- Calibración del modelo.....	125
1.2.2.- Evaluación bayesiana de la exactitud diagnóstica	126
1.2.2.1.- Capacidad discriminante.....	126
a) Estudio bayesiano de la curva ROC.....	126
b) Índices bayesianos de exactitud diagnóstica.....	128
1.2.2.2.- Calibración del modelo.....	135
a) Usando como estimador puntual por estrato el estimador bayesiano convencional.....	136
b) Usando como estimador puntual por estrato el estimador intrínseco.....	137
2.- 2ª Fase: validación de ambos tests.....	138
2.1- Modelo predictivo de Regresión Logística.....	138
2.1.1.- Evaluación clásica de la exactitud diagnóstica.....	138
2.1.1.1.- Capacidad discriminante.....	138
a) Estudio de la curva ROC.....	138
b) Índices de exactitud diagnóstica.....	140
2.1.1.2.- Calibración del modelo.....	140
2.1.2.- Evaluación bayesiana de la exactitud diagnóstica	141
2.1.2.1.- Capacidad discriminante.....	141

a) Estudio bayesiano de la curva ROC.....	141
b) Índices bayesianos de exactitud diagnóstica.....	143
2.1.2.2.- Calibración del modelo.....	150
a) Usando como estimador puntual por estrato el estimador bayesiano convencional.....	150
b) Usando como estimador puntual por estrato el estimador intrínseco.....	151
2.2- Modelo de Red Neuronal Artificial.....	151
2.2.1.- Evaluación clásica de la exactitud diagnóstica.....	154
2.2.1.1.- Capacidad discriminante.....	154
a) Estudio de la curva ROC.....	154
b) Índices de exactitud diagnóstica.....	155
2.2.1.2.- Calibración del modelo.....	155
2.2.2.- Evaluación bayesiana de la exactitud diagnóstica	156
2.2.2.1.- Capacidad discriminante.....	156
a) Estudio bayesiano de la curva ROC.....	156
b) Índices bayesianos de exactitud diagnóstica.....	158
2.2.2.2.- Calibración del modelo.....	165
a) Usando como estimador puntual por estrato el estimador bayesiano convencional.....	165
b) Usando como estimador puntual por estrato el estimador intrínseco.....	166
3.- 3ª Fase: comparación de la exactitud diagnóstica de ambos tests.....	167
3.1.- Comparación clásica de la diferencia en la exactitud diagnóstica.....	167
3.1.1.- Diferencia en la capacidad discriminante.....	167
3.1.1.1.- Análisis del área bajo las Curvas ROC (AUC).....	167
3.1.2.- Diferencia en la calibración.....	168
3.2.- Propuesta de evaluación bayesiana de la diferencia en la exactitud diagnóstica.....	169
3.2.1.- Diferencia en la capacidad discriminante.....	169
3.2.1.1.- Análisis bayesiano del área bajo las Curvas ROC (AUC).....	169
a) Utilizando la aproximación Normal.....	170
b) Utiliando la simulación con técnica de Monte Carlo.....	173
c) Utiliando el método gráfico discreto de la parrilla de Berry.....	174

3.2.2.- Diferencia en la calibración.....	175
3.3.- Comparación mediante la incorporación al proceso diagnóstico: análisis de decisión.....	176
3.3.1.- Evaluación clásica.....	178
3.1.2.- Evaluación bayesiana.....	182
VII. - Discusión.....	187
1.- La investigación científica sobre terapia: experimentos para "probar causalidad".....	188
1.1.- Versión frecuentista de la ciencia para probar causalidad.....	191
1.1.- Versión bayesiana de la ciencia para probar causalidad.....	195
2.- La investigación científica sobre diagnóstico: experimentos para "aprender".....	198
2.1.- Estudio de la capacidad discriminante.....	200
2.2.- Estudio de la calibración.....	210
2.3.- Análisis de decisión.....	214
3.- Consideraciones finales.....	214
VIII. - Conclusiones y proyección de futura investigación.....	217
IX. - Anexos.....	223
ANEXO I: Teorema de Ramsey-De Finnetti.....	225
ANEXO II: Intercambiabilidad y Teorema de Representación.....	231
ANEXO III: Test bayesiano de hipótesis: Razón de Verosimilitudes y Factor de Bayes.....	233
ANEXO IV: Teoría de la información.....	237
ANEXO V: El Índice "Pediatric Risk of Mortality" (PRISM).....	247
ANEXO VI: Redes neuronales artificiales.....	249
ANEXO VII: Uso clínico de un test diagnóstico: Modelo de Pauker-Kassirer modificado por Latour.....	271
X. - Bibliografía.....	279



Universitat d'Alacant
Universidad de Alicante



***II. - Justificación:
introducción y estado actual
del tema.***



Universitat d'Alacant
Universidad de Alicante

1.- LA PREDICCIÓN EN CIENCIA Y EL LLAMADO "PROBLEMA DE LA INDUCCIÓN"

El psicólogo James E. Alcock⁴ describe a nuestro cerebro como la "máquina de generar creencias", porque está constantemente procesando información procedente de nuestros sentidos y generando nuevas creencias sobre el mundo que nos rodea. El cerebro produce estas creencias de modo que sean coherentes con las previamente sustentadas pero sin preocuparse demasiado acerca de qué es verdad y qué no lo es, pues dispone de muy pocos medios para diferenciar las conexiones causales de las meras coincidencias⁵. Es su manera normal de funcionar, y ha evolucionado de esta manera para maximizar sus posibilidades de supervivencia.

El problema de aprender de los datos ha sido investigado por los filósofos a lo largo de toda la historia de la humanidad bajo el nombre técnico de inferencia inductiva. Comúnmente⁶, a la inducción también se le denomina creencia. La mayoría de las veces, cuando hablamos de inducción estamos pensando en inferir o predecir algo del futuro desde lo que conocemos en el presente. En 1739 y a la edad de 28 años, el filósofo escocés David Hume (1711-76) publicó *A Treatise of Human Nature*, y en él⁷ estableció la formulación clásica del llamado "problema lógico de la inducción", que estudió en profundidad en su *Enquiry Concerning the Human Understanding* publicado en 1748⁸. En esencia, el problema (que Bertrand Russell denominó "el escándalo de la filosofía"⁹) consiste en que *nunca* está justificada la inferencia mediante un razonamiento lógico que proceda por inducción: no existe ningún razonamiento inductivo que sea lógicamente válido. La inducción, de hecho, se asemeja al *modus operandi* del

"(...) señor que condujera un coche con el parabrisas pintado de negro y que intentara predecir el trazado de la carretera que tiene delante a base de mirar por el retrovisor: mientras la carretera es recta todo va bien, pero a la primera curva el coche se va directo a la cuneta"¹⁰.

Aunque aún hoy pueda parecernos sorprendente, los filósofos de la ciencia del siglo XX^{11,12} han establecido que la inducción pura, el aprendizaje o adquisición de nuevo conocimiento cierto derivado exclusivamente mediante especulación racional desde los datos relativos a hechos del pasado, es imposible a menos que se asuma como verdadero cierto conocimiento previo. En palabras de Sir Peter Medawar, premio Nobel de Medicina en 1960, "la inducción es un mito"¹³.

Es por ello que, en ciencia, hacer una predicción es una actividad arriesgada que se lleva a cabo en un contexto de total incertidumbre. Nunca tenemos la completa seguridad de que nuestra previsión va a cumplirse: siempre podemos errar en nuestros pronósticos¹⁴. Irónicamente, parafraseando a Xavier Sala i Martín, se podría decir que existen dos tipos de científicos: los que no saben hacer profecías (por lo menos profecías que siempre acaben cumpliéndose) y los que *no saben que no saben hacer profecías*¹⁰.

2.- LA TEORÍA DE LA PROBABILIDAD Y LA ESTADÍSTICA COMO VÍAS PARA SOLVENTAR EL PROBLEMA

El finlandés Georg Henrik Von Wright, profesor de lógica finlandés que fue alumno de Ludwig Wittgenstein y luego le sustituyó en la cátedra de filosofía de Cambridge, ha demostrado¹⁵ en su tesis de 1941 que la creencia en un juicio sintético (proposición lógica) cuya verdad se demuestre a priori es contradictoria, y por ello es imposible justificar lógicamente la inducción. Así, "*de las proposiciones que afirman la existencia de un objeto o la ocurrencia de un suceso, es imposible garantizar necesariamente proposiciones que afirmen la existencia de otro objeto o la ocurrencia de otro suceso diferentes de los primeros*". En palabras de Hume⁷: "*no hay ningún objeto que implique la existencia de otro cualquiera, si consideramos los objetos en sí mismos*". Ello prueba que el problema de la inducción no tiene solución: Hume estaba en lo cierto.

Pero en la época de Hume la teoría de la probabilidad acababa apenas de nacer. Así, parece que el filósofo escocés había leído la primera edición (1718) del libro *The Doctrine of Chances* de Abraham de Moivre¹⁶, descubridor de la aproximación normal a la distribución binomial^a, y no fue hasta 1763 cuando la Royal Society publicó el ensayo póstumo del Rvdo. Thomas Bayes con su famoso teorema¹⁷. El enorme desarrollo posterior de la teoría matemática de la probabilidad, axiomatizada definitivamente por el matemático ruso A. N. Kolmogorov en 1933¹⁸, y de la inferencia estadística moderna nos permitirán hacer mucho más de lo que Hume hubiera jamás imaginado. No sugeriremos que dichos desarrollos permitan solucionar el problema de la inducción, pero afirmamos con Ian Hacking¹⁹ que la estadística nos muestra dos caminos o maneras útiles de evitarlo: la evasión bayesiana y la frecuentista. La estadística bayesiana, evade el problema haciéndonos capaces de aprender deductiva y racionalmente de los datos que nos proporciona la experiencia. La estadística frecuentista, la estimación por intervalos de confianza y la teoría inferencial de Neyman-Pearson, aún reconociendo que no es posible la inferencia inductiva sobre casos particulares, nos permiten obtener justificación deductiva para afirmar con confianza la veracidad o falsedad de ciertas hipótesis causales relativas a grupos homogéneos de individuos, y así poder llevar a cabo comportamientos inductivos consecuentes.

Sin embargo, el problema no queda resuelto. Ambas evasiones sólo funcionan si se asume como cierto que el futuro será como el pasado. Mediante la estadística podremos inferir algo sobre el futuro desde lo que sabemos en el presente sólo, repitiendo a Hume^{6,7,8}, "*bajo la suposición de que el curso de la naturaleza continuará siempre uniformemente igual, de que existe conformidad entre el futuro y el pasado*". John Stuart Mill llama a este principio la **ley de uniformidad de la naturaleza**, y lo formula así en 1879²⁰: "*En la naturaleza se producen casos paralelos; lo que sucede una vez volverá a suceder, dado un grado suficiente de semejanza de las circunstancias*". Pero esta uniformidad es sólo una creencia

^a Publicada en su *Miscellanea Analytica* de 1738, es, de hecho, la primera demostración de un caso especial del Teorema del Límite Central.

nuestra, una costumbre, un hábito inductivo que -como tal- no tiene justificación racional. De todas maneras, nosotros somos así: si una regla ha sido siempre cierta en el pasado, nos parece razonable suponer con certeza que lo seguirá siendo en el futuro²¹. Aunque la probabilidad no soluciona el problema, de hecho no nos importa mucho, porque nos funciona bajo suposiciones fácilmente asumibles por compatibles con nuestra manera de ser.

3.- LA EVASIÓN FRECUENTISTA: COMPORTAMIENTO INDUCTIVO BASADO EN INFERENCIA DEDUCTIVA:

La evasión "frecuentista" del problema de la inducción se basa en la concepción frecuentista o clásica de la probabilidad, desarrollada sistemáticamente por el profesor de Harvard de origen austriaco Richard von Mises en 1928²² en base a las ideas recogidas por el matemático inglés John Venn en su libro *Lógica del azar* de 1866²³. Se sustenta matemáticamente en la llamada *Ley de los Grandes Números* (Teorema de Bernoulli^{b)} y toda la familia de Teoremas del Límite Central, deducibles desde los axiomas de la teoría de probabilidades y cuya idea principal es la propiedad de que las frecuencias relativas se estabilizan conforme el tamaño muestral crece indefinidamente. La probabilidad se define como el límite, cuando n tiende a ∞ , de esa frecuencia relativa.

$$\text{Prob} = \lim_{n \rightarrow \infty} r/n, \text{ siendo } r (r = \sum x_i) \text{ el número de pruebas positivas}$$

Este concepto de probabilidad no sólo tiene pues demostración matemática, sino que además se corresponde con un fenómeno objetivo y real que se puede experimentar empíricamente: una propiedad de la naturaleza.

La ciencia trata de determinar la verdad sobre cómo funciona la naturaleza. Pero aspira clásicamente a un conocimiento expresado en proposiciones del

^b El capítulo 5 de la parte IV del *Ars Conjectandi* de Jacques Bernoulli (1654-1705), publicado el 1713, demuestra el primer teorema del límite de la teoría de probabilidades.

lenguaje de la lógica formal; esto es, basada en el principio aristotélico del *tertium non datur*: las proposiciones lógicas sólo pueden ser verdaderas o falsas. Como hemos visto, para los frecuentistas la probabilidad es una propiedad que sólo puede aplicarse a una *serie* de eventos repetidos n veces: no tiene sentido hablar de probabilidad correspondiente a un evento o acontecimiento singular. Un evento aislado ocurre (en cuyo caso, una vez producido, su probabilidad individual de aparición es 1) o no ocurre (en cuyo caso es 0). Por ello, cuando se trata de aprender o adquirir conocimiento (generar hipótesis verdaderas sobre el comportamiento de la naturaleza) desde unos datos aisladamente observados, no tiene sentido hablar de la frecuencia (número relativo de veces) con que una hipótesis es verdadera. Las hipótesis, como todas las proposiciones de la lógica clásica, son necesariamente verdaderas o falsas y no a veces verdaderas y a veces no.

Para el conocimiento del mundo, los frecuentistas confirman la imposibilidad de una inferencia inductiva individual. Sin embargo, usando la teoría de probabilidades evitan el problema de la inducción porque utilizan un *método deductivo* para realizar inferencias y extraer conclusiones. La inferencia estadística frecuentista se vale de un método que es de carácter deductivo y diseñado para acertar una inmensa mayoría (usualmente el 95%) de las veces que se aplica. Por ello se habla de la frecuencia relativa con que las inferencias así realizadas son correctas: el llamado "grado de confianza" que nos da este método de realizar inferencias válidas (usualmente, por convención, el 95%). Así, si estamos interesados en conocer el valor verdadero (poblacional) de una cierta cantidad sobre la base de unos datos (muestrales) obtenidos por observación, se nos ofrece un método deductivo que nos dará como resultado estimado un intervalo de valores de esa cantidad. Y además somos capaces de estimar la frecuencia relativa con la que, si se sigue ese método, un intervalo así obtenido incluirá al verdadero valor buscado. Es el llamado método de estimación por intervalos de confianza, desarrollado técnicamente en los años 1930-40 por Jerzy Neyman, un matemático polaco formado en Londres y emigrado a los Estados Unidos²⁴.

Si, por el contrario, estamos interesados en conocer la verdad o falsedad de cierta hipótesis causal que intenta explicar el funcionamiento de la naturaleza, se utiliza un método de contraste de hipótesis desarrollado inicialmente por R. A. Fisher²⁵, pero estructurado matemáticamente en su versión definitiva también por Jerzy Neyman y por Egon S. Pearson en 1928^{26,27,28}. Nos parece que queda fuera del alcance de los objetivos de este estudio explicar en profundidad cómo funciona el método Neyman-Pearson de contraste de hipótesis en que se basa la estadística clásica para realizar inferencias causales sobre el funcionamiento de la naturaleza. Puede encontrarse en la extensa literatura estadística disponible en nuestro país y, de entre los libros disponibles en inglés, sobre este tema nos parecen especialmente recomendables los libros de Deborah Mayo²⁹, Ronald N. Giere³⁰, y el más reciente editado por Mark L Taper y Subhash R Lele³¹ en los que se describe en profundidad el entramado lógico que subyace al método.

Diremos únicamente que la estadística frecuentista considera que solo existe una manera científica de conocer los fenómenos causales en un mundo no determinista: realizar experimentos aleatorizados con los que contrastar nuestras hipótesis³². Para contrastar estas hipótesis con los fenómenos del mundo real se utiliza un *método deductivo* que tiene como esqueleto o armazón, y eso es lo que le da justificación lógica, el clásico argumento del *modus tollens* que ya describiera como válido Aristóteles (384-322 a.d.C.) en su *Organon*³³. Por ello el experimento es capaz de refutar una hipótesis, pero nunca de verificarla. Esta es la manera deductiva como los experimentos son capaces de solventar el problema de la inducción para conocer los fenómenos de la naturaleza desde los datos obtenidos en el mundo real: descartando como falsas las hipótesis incompatibles con los resultados de un experimento, y aceptando momentáneamente como verdaderas (hasta que otro experimento las refuta) las que no entran en contradicción con los datos. A pesar de que no abordó directamente en su obra el asunto del contraste estadístico de hipótesis en el sentido moderno³⁴, parece que fue Karl Popper el filósofo de la ciencia que comprendió el funcionamiento de este método^{11,35,36}: a base de conjeturas y refutaciones. El premio Nobel de Física de 1966, Richard P.

Feynman, en una famosa conferencia³⁷, nos lo explica con más detalle (la cursiva es nuestra):

"En general, para buscar una buena ley seguimos el proceso que detallaré a continuación. En primer lugar hacemos una suposición sobre dicha ley. Luego calculamos las consecuencias de dicha suposición para ver que implicaría esta ley si lo que hemos supuesto fuera correcto. A continuación comparamos los resultados del cálculo con lo que se produce en la naturaleza mediante un *experimento*; es decir lo comparamos directamente con lo que se observa, para ver si funciona. *Si no concuerda con el experimento, entonces es falso*. Es esta afirmación tan sencilla está la clave de la ciencia. No importa lo maravilloso que nos parezca aquello que hemos supuesto. Tampoco importa lo ingeniosos que seamos, ni quien realizó la suposición, ni como se llama el que la formuló, *si no concuerda con el experimento, es falso*. (...)

Ya se habrán dado cuenta de que por este método es posible intentar demostrar la falsedad de cualquier teoría. Si tenemos una teoría estructurada, una conjetura auténticamente científica, a partir de la cual puedan calcularse conjeturas que puedan contrastarse experimentalmente, es posible en principio acabar con ella. Existe siempre la posibilidad de demostrar la falsedad de una teoría bien formulada; pero observen que nunca podemos demostrar su veracidad. Supongamos que tienen ustedes una idea brillante, calculan las consecuencias y descubren una y otra vez que las consecuencias calculadas concuerdan con los experimentos. ¿Puede decirse que la teoría es verdadera?. No, simplemente no se ha demostrado que sea falsa. En el futuro podrían calcularse nuevas consecuencias que condujeran a nuevos experimentos que invalidaran la teoría. (...) Nunca estamos definitivamente en lo cierto: de lo único que podemos estar seguros es de estar equivocados. Sin embargo no deja de ser alentador que podamos tener teorías que duran tanto tiempo."

Una vez realizado el experimento aleatorizado, el método de Neyman-Pearson nos posibilita deductivamente a que rechacemos o aceptemos la hipótesis nula, y nos indica la probabilidad (frecuencia relativa de veces) de que siguiendo este método nos equivoquemos haciéndolo. A estas probabilidades se les llama, respectivamente, Errores tipo I y II. La inversa de esos errores, es la confianza con que el método nos permite refutar o afirmar una H_0 (en el caso de afirmarla, a esta confianza se le llama "poder" del estudio).

La [FIGURA 1], modificada del libro de J. A. Diez y C. Ulises Moulines³⁸, resume la versión moderna de todo el proceso. Como se observa, una vez decidido si se cumple la hipótesis nula, se deberá comprobar igualmente si se cumplen los supuestos auxiliares y las condiciones iniciales. Cuando se está evaluando una

relación causa-efecto, los supuestos auxiliares son los que quedan recogidos en las [TABLAS 1 y 2].

FIGURA 1: El contraste de Hipótesis en los experimentos aleatorios mediante el método inferencial de Neyman-Pearson. SA: supuestos auxiliares. CI: condiciones iniciales.

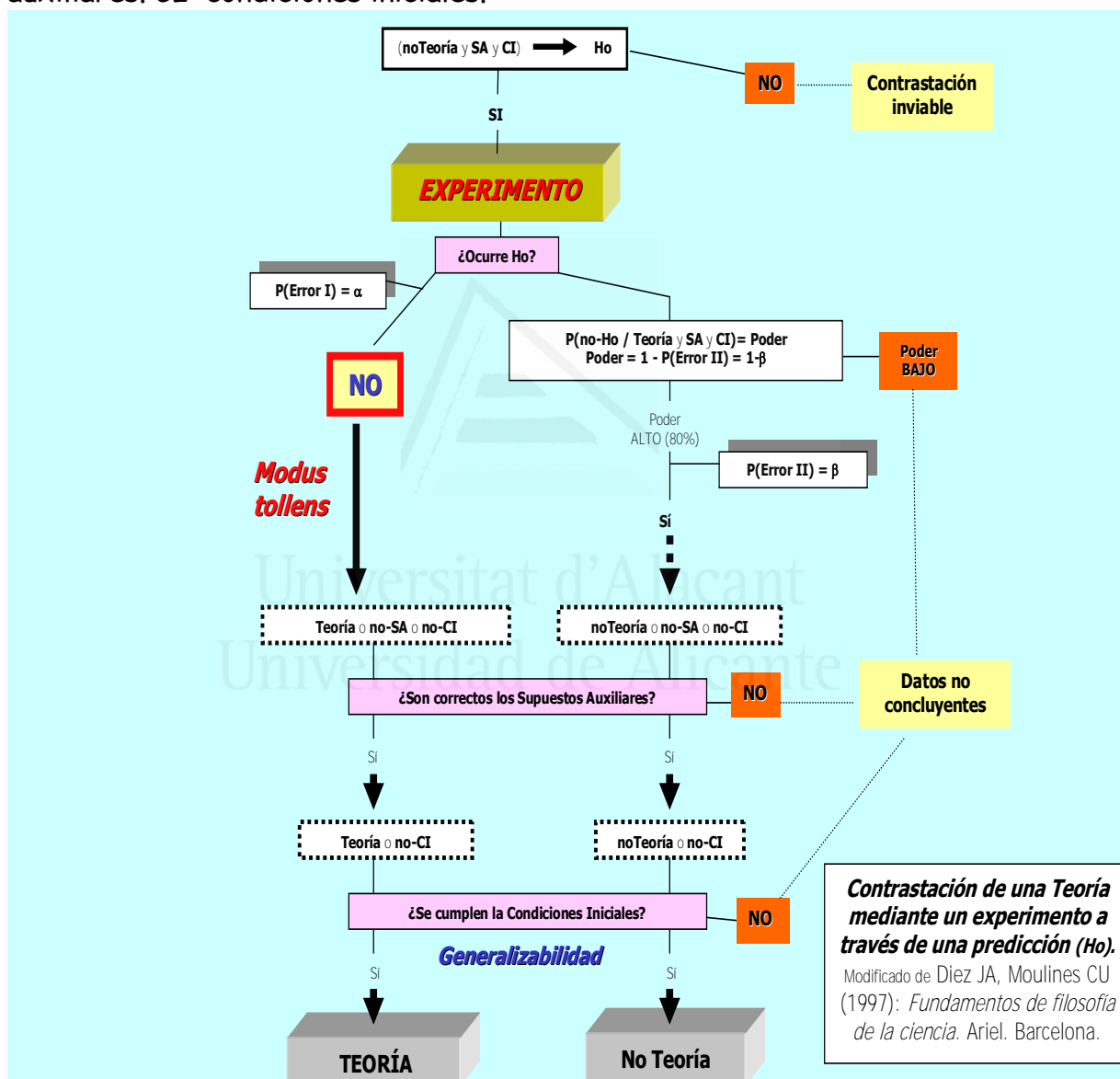


TABLA 1: Supuestos auxiliares para evaluar una relación causal (Modificado de Mayo DG²⁹)

1. - VARIABLES ADECUADAS Y CÁLCULOS BIEN HECHOS:

Uso de Computadora para los cálculos

Cumplimiento de condiciones de aplicación de los test estadísticos

2. - REQUISITOS DE EXPERIMENTO ALEATORIO para que pueda modelizarse por la estadística y la teoría de probabilidades

Distribución al azar pura (muestreo aleatorio)

No pérdidas en el seguimiento

Ocultación de la Secuencia de Aleatorización (OSA)

Análisis según criterio de "Intención de Tratar"

3. - REQUISITOS DE CAUSALIDAD

Sir A.B. Hill (1965) - U.S. Surgeon General (1964-89)

4. - AUSENCIA DE ERROR SISTEMÁTICO (SESGOS)

Aleatorización y Muestra Grande (n >30-50)

Control de la confusión e interacción (Anál. multivariante)

Enmascaramiento

Evitar sesgo de selección (muestra no representativa)

Evitar sesgo de información o de Clasif. Errónea

Evitar el sesgo de publicación

5. - COMPARABILIDAD DE LOS GRUPOS

Aleatorización y Muestra Grande (n >30-50)

Control de la confusión e interacción (Anál. multivariante)

TABLA 2: Requisitos de Causalidad (Modificado de Hill A.B³⁹, U.S. Surgeon General 1965: Smoking and Health⁴⁰, y U.S Surgeon General 1990: Criteria for evaluating evidence regarding the effectiveness of perinatal interventions⁴¹)

1.- Estudio: Diseño adecuado + Validez Interna

2.- Criterios mayores

- **Precedencia temporal correcta (E. Prospectivo):** Una intervención puede ser considerada causa de la reducción en el riesgo de una enfermedad/anormalidad perinatal si, y sólo si, aquella se aplicó antes del momento en el que se desarrolló ésta.
- **Plausibilidad biológica:** Un mecanismo biológicamente plausible debe ser capaz de explicar por qué tal relación podría darse el caso de ocurrir.
- **Consistencia:** Rara vez estudios únicos son definitivos. Resultados concordantes aportados por varios estudios diferentes realizados en poblaciones distintas y por investigadores diversos aportan más peso a la hipótesis de causalidad. Si los resultados de varios trabajos son inconsistentes, debería estudiarse primero la calidad metodológica de los estudios, pero no se excluye la causalidad.
- **Exclusión de explicaciones alternativas (Confusores):** La profundidad con la que se han estudiado posibles explicaciones alternativas es un criterio importante a la hora de juzgar la causalidad.

3.- Otros criterios

- **Gradiente dosis-respuesta:** Si un factor es efectivamente la causa de una enfermedad, normalmente (aunque no invariablemente) a mayor exposición al factor mayor riesgo de enfermedad. Tal relación dosis-respuesta no siempre aparece, porque algunas importantes relaciones biológicas son dicotómicas y se debe alcanzar cierto **nivel umbral** para observarse la respuesta. La ausencia de este gradiente no excluye causalidad, pero su presencia la refuerza.
- **Magnitud de la fuerza de asociación (RR/OR, DR) y Precisión de la estimación (IC estrecho)**
- **Efecto del cese de exposición:** A no ser que exista un efecto mantenido, cuando una intervención tiene un efecto beneficioso el beneficio debe desaparecer cuando aquella se elimina de la población.

Las condiciones iniciales en las que se ha realizado el experimento marcan la generalizabilidad de sus resultados: sólo podrán inferirse a una población que cumpla los requisitos (criterios de inclusión y exclusión) que describen la muestra aleatoria que le es representativa.

Ian Hacking^{19,42,43} nos ha mostrado que fue el filósofo americano Charles Sanders Peirce (1839-1914), quien estableció toda la base lógica sobre la que se fundamenta esta evasión del problema de la inducción. Hijo de un matemático de Harvard, fue el fundador de la corriente filosófica llamada "pragmatismo", aunque se ganaba la vida como científico empleado del gobierno en el Servicio Costero de Mediciones Geodésicas (su trabajo consistía en hacer mediciones y mejorar los artefactos de medir, por lo que estaba tremendamente familiarizado con la llamada "curva de errores" que luego pasaría a llamarse distribución normal gaussiana). Peirce considera que es imposible la inducción individual, pero tiene una idea de la lógica que unifica la deducción y la inducción. Para él, un argumento lógico es deductivamente válido (conservador de verdad) si su conclusión es siempre verdad cuando las premisas son ciertas, y un argumento lógico es inductivamente bueno si es de tal forma que su conclusión es frecuentemente verdad cuando las premisas son ciertas. Esto es: un argumento es 95% bueno (nos da una confianza del 95%) si está construido de forma que su conclusión es verdadera (suponiendo que las premisas lo sean) el 95% de las veces que se usa. Es el concepto desarrollado matemáticamente por Fisher en su teoría de diseño de experimentos, y ulteriormente por Neyman y Pearson en su teoría de contraste de hipótesis y de inferencia por intervalos de confianza.

El método deductivo propugnado por los frecuentistas nos solventa el problema de la inducción porque, aunque no podemos creer siempre sus conclusiones, nos permite llevar a cabo una *conducta o comportamiento inductivo*⁴⁴ basándonos en un razonamiento deductivo, pues estamos deductivamente seguros de que es un método que acierta la mayoría (usualmente el 95%) de las veces que se utiliza. Si nos comportamos de acuerdo con lo que el método nos ha hecho aprender de los datos, estaremos haciendo lo correcto la mayoría de las veces.

Esto es, según los frequentistas, todo lo que necesitamos para una ciencia objetiva y racional.

3.1.- USO DE LA EVASIÓN FRECUENTISTA EN LA MEDICINA CLÍNICA:

En medicina clínica, la realización de experimentos aleatorizados y la estimación inferencial por intervalos de confianza se utiliza desde 1948 en que se publicó, dirigido por Sir A. Bradford Hill, el primer ensayo clínico⁴⁵. El ensayo clínico es, hoy por hoy, el experimento de la medicina clínica: lo que le otorga carácter de ciencia y le da la capacidad para establecer conocimiento sobre relaciones causales. La aleatorización de muestras grandes, por su capacidad para minimizar el sesgo de confusión, es, para los frequentistas y en virtud de la Ley de Grandes Números, la esencia de la experimentación: el momento del diseño de un experimento en el que queda justificada lógicamente la validez de una inferencia causal.

Universitat d'Alacant
Universidad de Alicante

4.- LA EVASIÓN BAYESIANA: APRENDIZAJE RACIONAL DESDE LOS DATOS DE LA EXPERIENCIA:

La evasión "bayesiana" del problema de la inducción representa la manera racional de aprender de los datos de la experiencia, y resulta indispensable^{46,47} para tomar decisiones científicas de una manera coherente. Se fundamenta en la aplicación del Teorema de Bayes¹⁷, un resultado simple pero fundamental de la teoría de probabilidad clásica, aunque su desarrollo histórico es considerablemente moderno. Laplace y Gauss, hace ya doscientos años, trabajaron con distribuciones a posteriori con un método conocido con "el método inverso". Pero hablando

estrictamente, parece que la aplicación práctica de la estadística bayesiana se inició en 1959 con la publicación de libro de Robert Schlaifer *Probability and Statistics for Business Decisions*⁴⁸.

La metodología bayesiana surge de la confluencia de dos ramas de la matemática. Por un lado la teoría estadística de la decisión, basada en la teoría matemática de juegos (fundada por Von Neumann y Morgenstern⁹³ en 1942) y desarrollada por la obra de estadísticos como el mismo Jerzy Neyman⁴⁹ o Wald, y científicos aplicados como el físico sir Harold Jeffreys^{21,50}. Por otro, la concepción subjetivista del concepto de probabilidad, introducida ya en 1713 por Jakob Bernoulli (para quien la probabilidad era una magnitud determinada exclusivamente de manera subjetiva y personal, dependiente del conocimiento individual) y sugerida en 1924 por el matemático francés Emile Borel⁵¹, pero desarrollada sistemáticamente⁵² entre 1920 y 1940 desde la teoría general de la probabilidad independientemente por Frank Ramsey en Cambridge⁵³ y por Bruno de Finetti en Italia^{54,55,56}. La consolidación de ambas líneas se produce en 1954 con la publicación por Leonard J. Savage del libro *The foundations of statistics*⁹¹, en el que se desarrolla sólidamente toda su estructura lógica utilizando el método de deducción matemática desde una base axiomática. Más modernamente, y como base del llamado "análisis de referencia" desarrollado, entre otros, en la Universitat de València por el Prof. José M. Bernardo⁵⁷ para evitar el cariz "subjetivo" de las conclusiones de la inferencia bayesiana, se utilizan también conceptos de la teoría matemática de la información desarrollada por Claude E. Shannon^{58,59} en 1948.

4.1.- LA ESTADÍSTICA BAYESIANA COMO ESTUDIO DEL MANEJO RACIONAL DE LA INCERTIDUMBRE:

En el estudio de los fenómenos aleatorios y en la mayoría de los problemas en los que se tiene que tomar una decisión, aparece de forma natural la incertidumbre. Pero también muchas situaciones puramente deterministas⁶⁰ (p.ej.

¿Cuál es la probabilidad de que el dígito que ocupa la posición p de la parte decimal de π sea un número par?) presentan incertidumbre, porque no tenemos información suficiente o porque no disponemos de medios de cálculo adecuados que suplan nuestra incapacidad para procesar los datos. Determinar la mejor de un conjunto de alternativas sería, en principio, inmediato si tuviéramos información completa de cada una de ellas. Así: el vendedor de periódicos que debe decidir sobre el número de ejemplares con los que se queda no tendría ningún problema si supiese el número exacto que conseguirá vender; o el meteorólogo no se equivocaría en sus predicciones si el problema no fuera caótico, es decir, extremadamente sensible a las condiciones iniciales. Lo que observamos, y podemos medir, es sólo una posibilidad entre muchas, pero necesitamos saber cuál de todas ellas será la que realmente se producirá, y ello nos genera incertidumbre. La principal dificultad con que nos encontramos ante situaciones inciertas es, pues, la falta de información sobre lo que finalmente sucederá⁶¹. La reacción natural ante esa incertidumbre es eliminarla en lo posible, obteniendo e incorporando más información mediante un proceso efectivo de aprendizaje: obtener nuevos datos que puedan proporcionar información relevante y combinar esta información con aquella de que inicialmente se disponía para tomar entonces la decisión más apropiada.

Para los bayesianos, la estadística es el estudio de esa incertidumbre y de lo que se debe hacer para minimizarla efectivamente. El papel principal de la estadística es asistir a los individuos de otras disciplinas (meteorólogos, economistas, técnicos en agricultura, juristas, médicos...) en las que existe variabilidad en los datos que utilizan en sus campos de trabajo, y ello les produce un ambiente de incertidumbre. Mediante la estadística, todos estos "clientes" son capaces de manejar racionalmente la incertidumbre para tomar coherente y racionalmente sus decisiones.

Pero la estadística bayesiana, y en concreto lo que se conoce como moderna *Teoría General de la Decisión Estadística*, no pretende ser una **descripción** del comportamiento real de las personas que se enfrentan a una decisión en presencia de incertidumbre: multitud de experimentos han demostrado que, incluso en

situaciones simples, el comportamiento humano muestra muchas incoherencias. Pretende ser una teoría **normativa** de la decisión: busca el proceso lógico que debe seguirse, lo que se debe hacer para tomar una decisión racional. Para su desarrollo ha sido necesario seguir el mismo proceso que se ha aplicado a otras disciplinas para convertirlas en ramas de la ciencia matemática sólidamente fundadas: encontrar unos axiomas o principios básicos aceptados intuitivamente como verdades en sí mismas (los axiomas que definen la coherencia), y construir luego, mediante el poder de la lógica deductiva (conservadora de la verdad) y basándose en la teoría matemática de la probabilidad, todo un edificio teórico robusto para ayudar a optimizar la toma de decisiones en ambientes de incertidumbre.

4.2.- LA PROBABILIDAD COMO MEDIDA (SUBJETIVA) DE LA INCERTIDUMBRE O GRADO DE CREENCIA RACIONAL:

Aceptar que la estadística es el estudio de la incertidumbre implica que es necesario investigar científicamente la incertidumbre como fenómeno⁶². Y el abordaje científico de cualquier fenómeno implica su medición, pues, siguiendo a lord Kelvin, "*es sólo asociando números con cualquier concepto científico como ese concepto puede ser verdaderamente comprendido*". Y *¿cómo se consigue una medida?*, *¿cómo se lleva a cabo una medición?* Lo esencial cuando vamos a hacer una medición es comprender que **no existen absolutos** en el mundo de las medidas, que cualquier medida está basada en la comparación con un patrón de referencia. Así, p. ej. la distancia se describe en referencia a la longitud de onda de la línea naranja del espectro de la luz que emite el isótopo kriptón-86, y el tiempo en referencia a la oscilación de un cristal. Es, pues, esencial encontrar un patrón de referencia para poder medir nuestra incertidumbre sobre un evento: el grado de creencia (racional) en que finalmente acontecerá. Se han sugerido varios, pero quizá los más simples fueron también históricamente los primeros: los juegos de azar.

Tomaremos como ejemplo el propuesto por Lindley⁶². Sea una urna que contiene un número N de bolas físicamente idénticas (salvo en el color). De ellas, un número R de bolas son rojas y el resto blancas. Supongamos también que usted extrae de la urna una bola "al azar". Es decir, que usted piensa que el mecanismo de extracción de la bola es aleatorio: que si le ofrecieran un premio si saliera la bola nº 57 o el mismo premio si saliera la bola nº 12 o cualquiera de las otras N bolas, usted, **subjetivamente**, no preferiría ninguna de estas situaciones (sería indiferente a ellas). Pues bien, sea un evento cuya aparición es para usted incierta. Usted está interesado/a en medir esa incertidumbre subjetiva: el grado de creencia que usted tiene en que el evento finalmente se producirá. Si usted piensa que el evento es tan incierto como la extracción, mediante un mecanismo que usted estima aleatorio, de una bola roja de una urna que contiene N bolas de las cuales R son rojas, entonces el evento tiene incertidumbre R/N para usted. Para medir la incertidumbre podemos utilizar las probabilidades o proporciones en un grupo.

Obsérvese que, para encontrar esta medida hay una asunción que no es necesario hacer: no es preciso que la comparación con el patrón estándar se realice física y realmente, sino sólo que sea posible y razonable hacerlo, que podríamos hacerlo si supiéramos cómo. Que no sepamos como se puede realizar la medición exacta de la distancia de la tierra a la luna no implica que neguemos la existencia de tal distancia. No es preciso que muramos por ingestión de *Amanita phalloides* para que comprendamos que es una seta venenosa.

Pero la razón para buscar una forma de medir nuestra incertidumbre (o nuestro grado de creencia) no es sólo hacer más precisa la noción de que estamos más seguros de que ocurrirá un suceso que otro, sino -esencialmente- que seamos capaces de **combinar** tales incertidumbres. El uso de las probabilidades o proporciones en un grupo para medir incertidumbres, cobra especial interés en la medida en que nos permite tener reglas para combinarlas:

- a) una regla de **adición** para incertidumbres de eventos mutuamente excluyentes: supongamos que de las N bolas de la urna, R son rojas, A son azules y el resto son blancas. Sean dos eventos mutuamente

excluyentes cuyas medidas de incertidumbre son, respectivamente, R/N y A/N . Usar esas proporciones para medir su incertidumbre nos posibilita, según el cálculo de probabilidades, a decir que la incertidumbre asociada con el evento suma de esos dos eventos mutuamente excluyentes es igual a la incertidumbre asociada con la extracción aleatoria de una bola coloreada, esto es $(R + A) / N$.

b) una regla de *multiplicación* de incertidumbres: supongamos que de las N bolas de la urna, R son rojas, y las restantes $(N - R)$ son blancas. Al mismo tiempo L son lisas y $(N - L)$ no lo son (son rugosas como las de golf). Así, la urna contiene cuatro tipos de bolas, y T es el número de bolas que son rojas y lisas. Usar proporciones para medir incertidumbres nos posibilita también, usando de nuevo el cálculo de probabilidades, a decir que la incertidumbre asociada con el evento cuya incertidumbre sea T/N es exactamente igual al producto de las incertidumbres asociadas con los eventos cuyas incertidumbres sean R/N y L/N , esto es $R/N \times L/N$.

c) una regla de *convexidad* de incertidumbres: la certeza de que las medidas de incertidumbre siempre pertenecen al intervalo (convexo) unidad $[0,1]$.

Pues bien, no sólo sabemos que **podemos** usar las probabilidades o proporciones. La estadística bayesiana hace uso del llamado^{52,63} *Teorema de Ramsey-de Finetti* para concluir deductivamente que **debemos** usar las probabilidades o proporciones para medir incertidumbres (grados de creencia): la única manera de evitar incoherencias en la medición de los grados de creencia es utilizar como medidas las probabilidades. *Las medidas de incertidumbre subjetiva deben obedecer las reglas del cálculo de probabilidad.* En el [ANEXO I] se especifica una prueba matemática de dicho teorema. La idea es que debemos estar seguros de que expresamos nuestros grados de creencia por medio de probabilidades, y así evitaremos la incoherencia que significa participar en una apuesta cuyo resultado es que siempre, hagamos lo que hagamos, perdemos dinero.

La única manera satisfactoria de expresar nuestras incertidumbres es con probabilidades: igual que un arquitecto sabe que para que la casa que concibe aguante en pie una vez construida, los cálculos numéricos que hace deben seguir las reglas de la aritmética, un decidor coherente debe saber que para tomar una decisión adecuada, los grados de creencia deben expresarse como probabilidades.

La concepción subjetivista del concepto de probabilidad interpreta, pues, a esta como **una medida** de la incertidumbre o el grado de creencia (racional) sobre la aparición de un determinado suceso, basada siempre en los datos previos de observación o experiencia. Ello se corresponde con el uso semántico convencional de esa palabra en el lenguaje ordinario cotidiano, en frases como "es poco probable que el hombre llegue a Marte antes de 2015", "la proporción de seropositivos en València es probablemente menor de 0'5% (en concreto la probabilidad de que esta proporción sea menor de 0'5% es de 0'84)", o "la probabilidad de que llueva durante la ceremonia de canonización de Juan Pablo II es del 80%". Este concepto intuitivo, habitual y cercano, no requiere de simetrías (como en el concepto clásico de probabilidad, basado en casos favorables y casos posibles) ni de la existencia de posibles repeticiones del mismo fenómeno (como en el concepto frecuentista convencional, basado en frecuencias relativas), pero incluye como casos particulares a los conceptos clásico y frecuencialista⁴⁶. La formulación matemática de la estadística bayesiana utiliza el concepto de probabilidad de un suceso como algo necesariamente subjetivo: como el grado de creencia (racional) en, o la incertidumbre sobre, su ocurrencia percibido por el sujeto en el momento que la expresa.

Esta interpretación del concepto de probabilidad fue desarrollada independientemente por Ramsey y De Finetti. Para ambos, no existen probabilidades absolutas: todas las probabilidades son **condicionadas** a un cierto estado de información previo sobre las condiciones en las que se lleva a cabo el suceso, y por tanto son siempre subjetivas. La probabilidad no tiene existencia objetiva: el fenómeno de estabilidad de las frecuencias está condicionado a que se cumplan las condiciones de experimento aleatorio. La probabilidad no es, pues, una

propiedad de la naturaleza, pero tampoco de los individuos: describe una relación entre un individuo y la naturaleza. $P_i(E/H)$ es una medida del grado de creencia en la incertidumbre sobre la ocurrencia del suceso E , asignada por un individuo i , en unas condiciones H .

La notación convencional para las probabilidades $P(E/H)$ en lugar de $P_i(E/H)$, omite la referencia al sujeto i que las especifica. Pero no existe ningún argumento que permita demostrar matemáticamente que dos individuos con la misma información deben, necesariamente, asignar las mismas probabilidades a sucesos inciertos. Las probabilidades lógicas en el sentido de Keynes⁶⁴ o de Carnap⁶⁵ no están matemáticamente justificadas. Esta actitud aparentemente solipsista es, sin embargo, compatible con la existencia de un amplio conjunto de probabilidades sobre cuyo valor puede existir consenso. Especialmente en el caso del diseño y análisis de la experimentación científica es posible elaborar probabilidades de consenso. Como veremos más adelante, para los bayesianos el desarrollo científico se apoya precisamente en la existencia de esas probabilidades razonable y coherentemente intersubjetivas. Así, mientras no tengamos posibilidad de comprobarlo, asumimos como científica la siguiente aseveración: la probabilidad de que el dígito 10^6 del número π sea 0 es 0'1.

4.3.- LA EVASIÓN BAYESIANA DEL PROBLEMA DE LA INDUCCIÓN:

Ya hemos visto (pág. 6) cómo la ciencia trata de determinar la verdad sobre cómo funciona la naturaleza, y aspira a un conocimiento expresado en proposiciones del lenguaje de la lógica formal clásica (*tercio excluso o principio de ambivalencia*): las proposiciones lógicas sólo pueden ser verdaderas o falsas. Tal y como decíamos que sostienen los frequentistas, pero a diferencia de la moderna "lógica borrosa" - para la que todo es cuestión de grado^{66,67} y por tanto el grado de verdad de un enunciado puede representarse por un número real entre 0 y 1 (su función de

pertenencia o inclusión al conjunto de cosas verdaderas)⁶⁸-, el paradigma bayesiano también afirma que las hipótesis sólo pueden ser verdaderas o falsas. Pero para éste la probabilidad es una medida condicional, en una escala [0,1], de la incertidumbre sobre -o el grado de creencia (racional) asociado a- el hecho de que una determinada hipótesis resulte finalmente ser verdadera. Los valores extremos 0 y 1, típicamente inaccesibles, describen respectivamente la imposibilidad o la certeza subjetivas en la verdad de una hipótesis u ocurrencia de un evento. La probabilidad no mide grados de verdad, sino *nuestra incertidumbre* sobre la veracidad de dichas hipótesis.

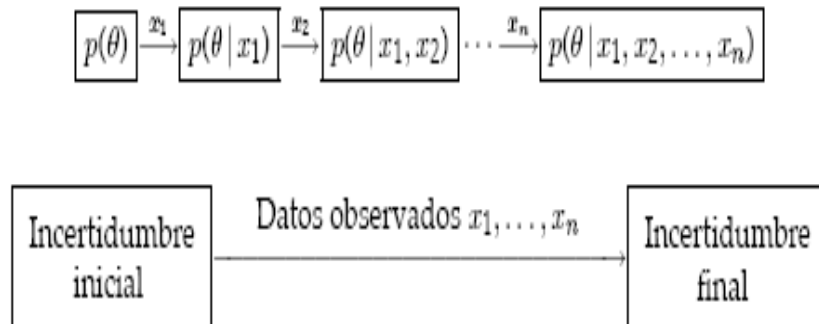
Una vez hemos expresado nuestra incertidumbre sobre un evento ¿qué hacemos con esa medida?, ¿cómo manejamos esa incertidumbre para obtener nuevo conocimiento? En definitiva, ¿cómo solventa el problema de la inducción la evasión bayesiana? La estadística bayesiana sólo necesita de la interpretación "subjetiva" de la probabilidad y de las matemáticas del cálculo de probabilidades. Partiendo de unos axiomas (enunciados que se asumen como verdaderos por sí mismos), es capaz de deducir matemáticamente todos sus principios y métodos. Los problemas de inferencia estadística se reducen así a problemas de teoría de probabilidades: el fundamento de la inferencia es un *método deductivo* que da validez lógica al conocimiento inductivo que obtiene. La inferencia estadística bayesiana está firmemente ("la fuerza de la deducción, que es conservadora de la verdad") basada en la verdad de sus fundamentos axiomáticos; la deducción matemática le proporciona una estructura lógica unificada, y garantiza la consistencia mutua de sus métodos.

Para los bayesianos, la reacción natural de cualquiera que deba tomar una decisión en presencia de sucesos o hipótesis inciertas, es eliminar cuanta incertidumbre le sea posible obteniendo más información sobre el problema. El decisor empieza por precisar la información de que inicialmente dispone sobre las incertidumbres que afectan a la situación; procura entonces obtener nuevos datos que puedan proporcionarle información relevante y, finalmente, combina esta nueva información con aquella de que inicialmente disponía para tomar entonces la

decisión más apropiada. La forma *racional* (coherente o no contradictoria) de llevar a cabo este proceso natural es, para los bayesianos, utilizar el Teorema de Bayes¹⁷. Así, so pena de ser incoherentes o contradictorios, se demuestra que debemos expresar la información inicial sobre la incertidumbre de la que partimos en forma de grados (racionales) de creencia representados por números con las características matemáticas de probabilidades (que cumplan los axiomas de Kolmogorov¹⁸ en la forma propuesta por Ramsey y De Finetti: convexidad, ley de adición de sucesos excluyentes, y ley de multiplicación [**ANEXO I**]). Para reducir la incertidumbre, se obtienen nuevos datos y la información que éstos proporcionan se describe mediante la función de verosimilitud. Ahora, la información inicial, descrita por probabilidades iniciales, es combinada con la información aportada por los nuevos datos obtenidos, descrita por la función de verosimilitud, y se obtienen las probabilidades finales: el grado de incertidumbre asociado a la información final total de que se dispone sobre el proceso. Para ello se utiliza el Teorema de Bayes, un resultado básico de la teoría matemática de probabilidades. Este proceso por el que, de una manera racional se ponen al día, de un modo internamente consistente, grados de creencia a la luz de nuevas evidencias, es el denominado *proceso de aprendizaje* que proponen los bayesianos para evadir el problema de la inducción: su propuesta para cambiar racionalmente nuestras creencias a la luz de las evidencias.

Básicamente, un experimento estadístico es, para los bayesianos, un proceso de reducción o disminución de la incertidumbre. Si el experimento es concluyente, aunque ello sólo ocurra de manera asintótica, la incertidumbre se reduce totalmente y es posible llegar a la certeza sobre las hipótesis partiendo desde nuestra creencia inicial. El diagrama siguiente⁶⁰ [**FIGURA 2**] resume cómo se realiza el proceso de aprendizaje. Se trata de una reducción secuencial de la incertidumbre inicial, $p(\theta)$, a medida que se van observando los datos $x_1, x_2, x_3, \dots, x_n$, mediante la aplicación secuencial del Teorema de Bayes:

FIGURA 2: El proceso de aprendizaje para reducir la incertidumbre



La incertidumbre residual sobre el parámetro θ en la etapa i -ésima, una vez observados los datos $x_1, x_2, x_3, \dots, x_{i-1}, x_i$, viene proporcionada por la **densidad a posteriori** $p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i)$ que se calcula mediante el teorema de Bayes del cálculo de probabilidades:

$$p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i) = \frac{p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1}) p(\theta / x_1, x_2, x_3, \dots, x_{i-1})}{\int p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1}) p(\theta / x_1, x_2, x_3, \dots, x_{i-1}) d\theta}$$

El cálculo se simplifica mucho usando la fórmula:

$p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i) \propto p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1}) p(\theta / x_1, x_2, x_3, \dots, x_{i-1})$
 pues la constante de proporcionalidad $[1 / \int p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1}) p(\theta / x_1, x_2, x_3, \dots, x_{i-1}) d\theta]$ se determina fácilmente con la condición de que la **densidad a posteriori** $p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i)$ debe ser una densidad de probabilidad, es decir integra la unidad con respecto al parámetro θ ($\int p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i) d\theta = 1$). La teoría de la probabilidad también nos indica que, si las observaciones son (condicionalmente) independientes, el resultado anterior se simplifica aún muchísimo más:

$$p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i) \propto p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1}) p(\theta / x_{i-1})$$

El grado de nuestra incertidumbre sobre el parámetro θ tras haber observado los datos $x_1, x_2, x_3, \dots, x_{i-1}, x_i$ de la muestra una vez realizado el i -ésimo experimento

estadístico, viene descrita por la densidad a posteriori $p(\theta / x_1, x_2, x_3, \dots, x_{i-1}, x_i)$. Se llama "principio de verosimilitud" al hecho de que toda la información que proporciona el experimento i -ésimo, entra al proceso en cada paso codificada mediante $p(x_i / \theta, x_1, x_2, x_3, \dots, x_{i-1})$, que no es una probabilidad de θ -aunque lo es de x_i - por lo que recibe otro nombre: la llamada *función de verosimilitud de θ* según los datos. Es por ello que, en inglés, todo el proceso se retrata con una famosa "frase hecha": *posterior is prior times likelihood*⁶⁹.

En la [FIGURA 3] se representa gráficamente un ejemplo de este proceso de aprendizaje. Los datos son ficticios, aunque están modificados -en aras de la claridad- de los publicados en el estudio GREAT^{70,71,72}: un ensayo clínico aleatorizado y controlado con placebo para demostrar la efectividad de la trombolisis precoz con anistreplase tras un infarto agudo de miocardio (IAM), administrada por los médicos de familia en casa del enfermo mientras llega el SAMU. Se utiliza como variable dependiente principal la mortalidad al mes del IAM, y el efecto beneficioso del nuevo tratamiento experimental se mide usando la Odds Ratio (OR), esto es, la razón de la Odds de morir en los asignados al nuevo tratamiento dividida por la Odds de morir al mes en los asignados a placebo [OR < 1 favorece al nuevo tratamiento]. Las distribuciones son todas normales en una escala de la variable transformada $\theta = \ln(OR)$.

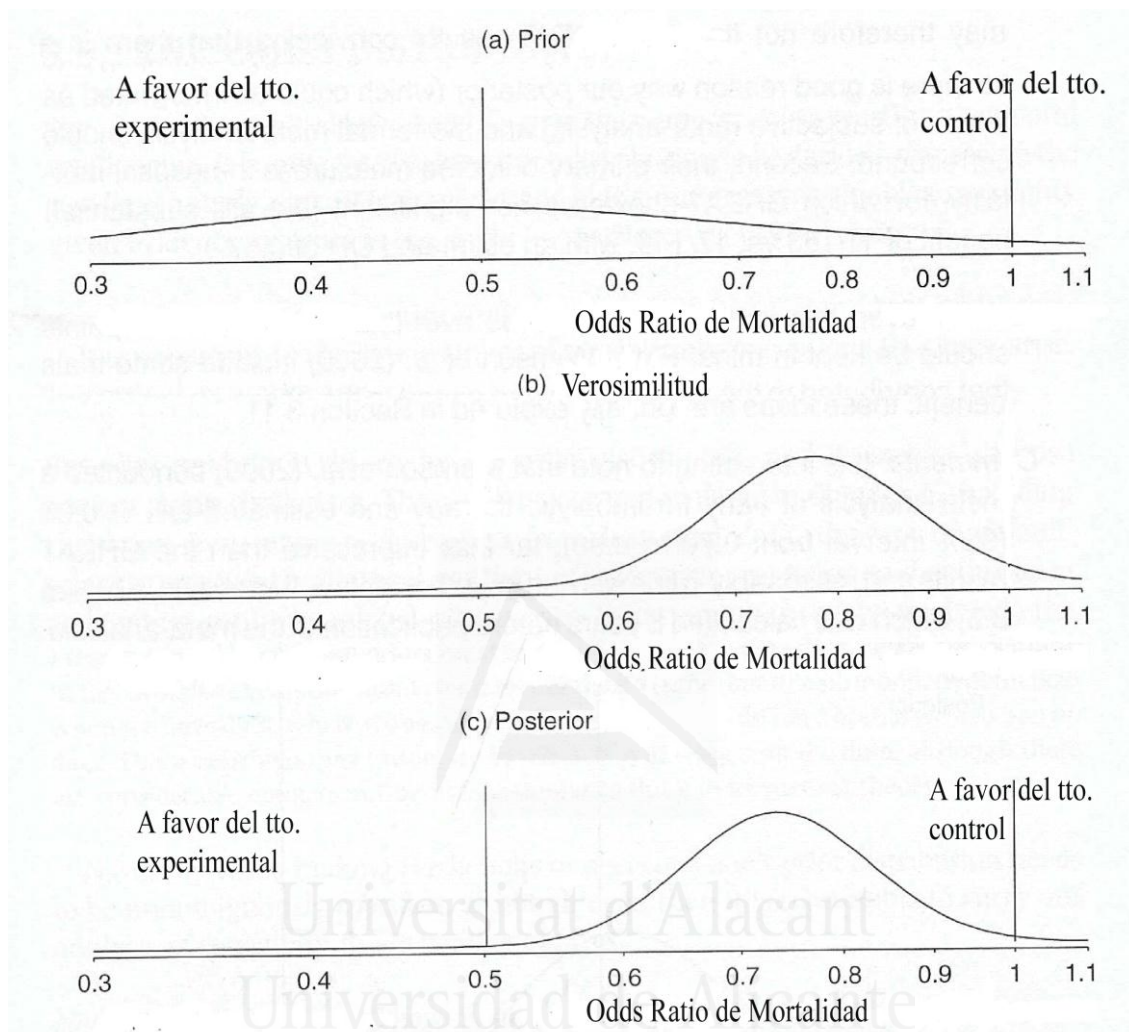


FIGURA 3: La información previa a la realización de un experimento nos hace representar la incertidumbre sobre el OR de mortalidad con una distribución de $\ln(\text{OR}) = N[-0.74, 0.131]$: favorable al tratamiento experimental, pero muy poco precisa. Se realiza un experimento, y los resultados de este coinciden con una función de verosimilitud representada por la curva normal $\ln(\text{OR}) = N[-0.26, 0.017]$. Utilizando el Teorema de Bayes actualizamos nuestra creencia (racional) previa aprendiendo de los datos del experimento. La curva resultante de nuestro aprendizaje sobre la incertidumbre del $\ln(\text{OR})$ es $N[-0.31, 0.0144]$. De nuestra incertidumbre previa, hemos pasado a estimar en un 95% la probabilidad de que la OR se sitúe entre 0.58 y 0.93. La probabilidad de que el tratamiento experimental sea efectivo ($\text{OR} < 1$) es de 0.995.

La probabilidad "a priori" que se ha escogido representa el juicio subjetivo de un especialista en cardiología que, conociendo y analizando críticamente las evidencias disponibles hasta esa fecha en la literatura especializada, estimó que era posible que el tratamiento produjera una reducción de la mortalidad, aunque consideró muy difícil precisar la cuantía de la OR: era posible cualquier valor entre 0.2 y 0.9, aunque parecía "a priori" algo más probable una reducción a la mitad. Se usó pues, para la distribución a priori de θ , una curva Normal con media = - 0.74 [OR = 0.48] y varianza = 0.13. Tras realizar el ensayo aleatorizado, el resultado fue de OR = 0.75 (IC95% de 0.61 a 0.99, $p = 0.04$) que, transformado a una verosimilitud normal de θ , coincide con una curva Normal con media = - 0.26 y varianza = 0.017. Usando el teorema de Bayes mediante el llamado "análisis conjugado normal", se multiplica la distribución previa por la verosimilitud obtenida en el ensayo, y se busca la constante de proporcionalidad para que el área total bajo la curva de la distribución "a posteriori" integre a uno. Así resulta que la distribución posterior de θ es una curva Normal con media = - 0.31 [OR = 0.73] y varianza = 0.0144. Esto coincide con un (bayesiano) intervalo de credibilidad al 95% de - 0.55 [OR = 0.58] a -0.07 [OR = 0.93]. Con ello hemos reducido notablemente nuestra incertidumbre: una vez incorporados los resultados del experimento, hay un 95% de probabilidad de (entendida aquí como grado de creencia racional en) que la OR se sitúe entre 0.58 y 0.93. Además podemos calcular que la probabilidad de que el tratamiento sea efectivo ($OR < 1$) es de 0.995.

4.3.1.- El teorema de Bayes para eventos simples⁷³:

De las reglas 1 a 3 del [ANEXO I] se derivan varias propiedades sencillas. Ya que $p(E \& H) = p(H \& E)$, la regla 3 implica que $p(H/E)p(E) = p(E/H)p(H)$, o lo que es lo mismo

$$P(H/E) = [p(H)/p(E)]p(E/H)$$

iiEsta es la demostración del Teorema de Bayes para el caso simple!! Como hemos visto este básico pero vital resultado de la teoría de probabilidades nos explica

cómo una probabilidad inicial $p(H)$ se convierte en una probabilidad condicional $p(H/E)$ cuando se considera lo aprendido de la ocurrencia del evento E , que se incorpora al análisis mediante la verosimilitud $p(E/H)$.

Consideremos ahora el caso de dos hipótesis alternativas H_0 y H_1 , mutuamente exhaustivas y excluyentes (una y sólo una de ellas es verdadera), capaces de explicar el evento observado E . La probabilidades "a priori" sobre cada una de ellas son $p(H_0)$ y $p(H_1)$. Supóngase que se ha observado un evento E , tal como el resultado de un test diagnóstico, y sabemos que las probabilidades de que tal evento ocurra siendo verdad H_0 y H_1 (verosimilitudes) son, respectivamente $p(E/H_0)$ y $p(E/H_1)$. El teorema de Bayes nos enseña a actualizar nuestras creencias a la luz de la evidencia, para producir las probabilidades posteriores. Así:

$$p(H_0/E) = [p(H_0)/p(E)]p(E/H_0) \quad \text{y} \quad p(H_1/E) = [p(H_1)/p(E)]p(E/H_1)$$

siendo $p(E) = p(E \& H_0) + p(E \& H_1) = p(E/H_0) p(H_0) + p(E/H_1) p(H_1)$.

Ahora, como $H_1 = \text{no}H_0$, entonces $p(H_1) = 1 - p(H_0)$ y $p(H_1/E) = 1 - p(H_0/E)$. Dividiendo las dos expresiones anteriores miembro a miembro:

$$p(H_0/E) / p(H_1/E) = [p(E/H_0) / p(E/H_1)] * [p(H_0) / p(H_1)]$$

En esta nueva expresión, $p(H_0) / p(H_1)$ es la ODDS "a priori" (un cociente de dos probabilidades complementarias) y $p(H_0/E) / p(H_1/E)$ es la ODDS "a posteriori", ambas a favor de H_0 . Es pues una expresión del teorema de Bayes en términos de ODDS. El cociente $p(E/H_0) / p(E/H_1)$ es la razón de las dos verosimilitudes, por lo que se le llama *cociente de verosimilitudes* (en inglés *likelihood ratio*) para el evento E , y también "Factor de Bayes" [ANEXO III]. Así que el teorema de Bayes puede expresarse igualmente como:

$$\text{Odds posterior} = \text{Odds prior} \times \text{Razón de Verosimilitudes}$$

Tomando logaritmos neperianos, observamos así mismo que:

$$\ln(\text{Odds post}) = \ln(\text{Odds prior}) + \ln(\text{likelihood ratio})$$

y al término $\ln(\text{likelihood ratio})$ se le denomina también⁷⁴ *peso de la evidencia*, nombre inventado por Alan Turing cuando utilizó estas técnicas para descubrir los códigos "Enigma" de las transmisiones nazis para los servicios de inteligencia aliados en Bletchey Park durante la segunda guerra mundial^{75,76}.

El cociente de verosimilitudes (factor de Bayes) transforma la Odds previa en la Odds posterior⁸⁴. Su valor varía entre 0 e ∞ , y sus valores pequeños se consideran tanto evidencia contra H_0 (el *numerador* del cociente) como evidencia a favor de H_1 (el *denominador* del cociente). Para interpretar su resultado, el físico bayesiano sir Harold Jeffreys diseñó la siguiente tabla [TABLA 3]:

TABLA 3: Calibración del Cociente de Verosimilitudes (Factor de Bayes)⁵⁰:

	RANGO DEL COCIENTE DE VEROSIMILITUDES (LR)	"FUERZA" DE LA EVIDENCIA
A FAVOR DEL <i>NUMERADOR</i> Y CONTRA EL DENOMINADOR	> 100	Decisiva
	32 a 100	Muy fuerte
	10 a 32	Fuerte
	3'2 a 10	Sustancial
	1 a 3'2	"Solo merece simple mención"
A FAVOR DEL <i>DENOMINADOR</i> Y CONTRA EL NUMERADOR	1 a 1/3'2 (1 a 0.3125)	"Solo merece simple mención"
	1/3'2 a 1/10 (0.3125 a 0.1)	Sustancial
	1/10 a 1/32 (0.1 a 0.03125)	Fuerte
	1/32 a 1/100 (0.03125 a 0.01)	Muy fuerte
	< 1/100 (< 0.01)	Decisiva

En un registro más informal⁷⁷, puede utilizarse la siguiente tabla [TABLA 4] para la interpretación del cociente de verosimilitudes (factor de Bayes):

TABLA 4: Interpretación informal del Cociente de Verosimilitudes⁷⁷:

	RANGO DEL COCIENTE DE VEROSIMILITUDES (LR)	"FUERZA" DE LA EVIDENCIA
A FAVOR DEL <i>NUMERADOR</i> (CONTRA EL DENOM)	≥ 10	Muy fuerte
	5	Fuerte
	< 2	Débil
A FAVOR DEL <i>DENOMINADOR</i> (CONTRA EL NUM)	> 1/2	Débil
	1/5	Fuerte
	$\leq 1/10$	Muy fuerte

4.3.2. - La predicción bayesiana

Una vez se ha aprendido de los datos del experimento, realizar predicciones de futuras observaciones es uno de los objetivos fundamentales de la actividad

científica, y la estadística bayesiana cumple con esta función de una manera muy elegante. Sean los datos ya observados en el experimento bayesiano un conjunto de observaciones intercambiables $\text{Datos} = x_1, x_2, x_3, \dots, x_n$, que pueden considerarse una muestra aleatoria de la distribución a posteriori $p(\theta / \text{Datos} = x_1, x_2, x_3, \dots, x_n)$, y supóngase que se quiere predecir el valor y de una observación futura generada por el mismo mecanismo aleatorio que ha generado los datos. Para la estadística bayesiana, la solución a este problema de predicción consiste en especificar la llamada *distribución predictiva* (posterior) $p(y/\text{Datos})$. Si la información contextual recabada nos indica que tanto los datos como la futura observación forman una secuencia intercambiable (no importa el orden sino sólo sus valores, véase más adelante pág. 56), de la teoría básica de la probabilidad (vease [ANEXO I]) se deduce por la ley de marginalización, "extendiendo la conversación" para incluir el parámetro θ , que:

$$p(y/\text{Datos}) = \int p(y/\text{Datos}, \theta) * p(\theta/\text{Datos}) d\theta$$

Como los $\text{Datos} = x_1, x_2, x_3, \dots, x_n$ y la nueva observación y son condicionalmente independientes dado θ , entonces $p(y/\text{Datos}, \theta) = p(y/\theta)$, por lo que la distribución predictiva queda

$$p(y/\text{Datos}) = \int p(y/\theta) * p(\theta/\text{Datos}) d\theta$$

que es un promedio de la verosimilitud de θ según el valor de la nueva observación utilizando como peso las creencias aprendidas sobre θ codificadas con su distribución a posteriori. Si se puede realizar esa integración, la predicción es muy sencilla. En el caso particular de los sucesos de Bernoulli, la distribución a posteriori es una Beta, y la distribución predictiva de un valor futuro $y=1$ es una Beta-Binomial bajo la cual podemos calcular un intervalo de predicción al 95%.

Si todas las asunciones (intercambiabilidad, modelo probabilístico...) son correctas, cuando el tamaño muestral de las nuevas observaciones aumente la distribución predictiva $p(y/\text{Datos})$ convergerá en la distribución a posteriori $p(\theta / \text{Datos})$ que generó los datos. De hecho, la mejor técnica para asegurar la calidad de las inferencias sobre θ codificadas mediante $p(\theta / \text{Datos})$ consiste en comprobar

con los datos observados la distribución predictiva $p(y / \text{Datos})$ generada por $p(\theta / \text{Datos})$.

4.4.-LA DISTRIBUCIÓN "A PRIORI" Y LA ILUSIÓN DE OBJETIVIDAD EN CIENCIA:

El uso repetido del teorema de Bayes como base de la inferencia constituye la esencia de la metodología bayesiana, el elemento que le da nombre. Pero la característica más importante que esta evasión de la inducción plantea, y el elemento que conceptualmente más problemas y rechazos genera (sobretudo entre los partidarios de la evasión frecuentista o clásica^{78,79}), es el uso del concepto "subjetivo" de probabilidad: la *necesidad matemática* de describir todas las incertidumbres presentes en los problemas de inferencia mediante distribuciones de probabilidad⁸⁰. Los parámetros desconocidos en los modelos probabilísticos *deben* tener siempre una distribución de probabilidad conjunta que describa toda la información disponible sobre sus valores en cada uno de los momentos del análisis, incluso al principio de todo el proceso de inferencia (densidad *a priori*). Es decir, a diferencia de lo que ocurre en la estadística frecuentista, los parámetros se consideran variables aleatorias: ello, sin embargo, no es una descripción de su variabilidad (los parámetros son realmente cantidades fijas desconocidas), sino una descripción de la incertidumbre sobre -grado de creencia en- sus auténticos valores⁸¹.

En las revistas científicas, se pretende que los análisis estadísticos doten de objetividad a las conclusiones de los estudios. Y el problema, origen de un larguísimo y fundamental debate entre las dos escuelas de estadísticos, está en que los bayesianos utilizan información "subjetiva" para iniciar un método de inferencia inductiva con el que pretenden alcanzar conclusiones "objetivas". Basan sus conclusiones en una combinación de: a) medidas "subjetivas" de la incertidumbre o grado de creencia sobre un conjunto de valores posibles del

parámetro de interés, codificadas por las distribuciones *a priori*, con b) datos "objetivos", extraídos de los experimentos y codificados mediante la función de verosimilitud. Ello parece atentar contra la aspiración de objetividad que subyace a toda actividad científica⁸², e incluso los bayesianos reconocen [FIGURA 4] que empíricamente puede comprobarse cómo existe una dependencia de las probabilidades a posteriori con respecto a la asignación (inicialmente subjetiva) de probabilidades a priori.

Desde la perspectiva bayesiana, sin embargo, esa objetividad es imposible en la ciencia. Los métodos estadísticos frecuentistas -valores de p bajo H_0 , intervalos de confianza-, que clásicamente se utilizan para revestir de validez objetiva a las conclusiones sobre las hipótesis científicas, son incapaces de proveer de esa objetividad. Usando los p -valores, las conclusiones se derivan no sólo de los datos que se han producido en el experimento, sino de todo el conjunto de valores más extremos que ellos (que, de hecho, no se han producido). Además, la significación estadística es dependiente del tamaño muestral: los mismos resultados experimentales son estadísticamente significativos (rechazo H_0) con muestras grandes y no lo son (no rechazar H_0) con muestras pequeñas. Un intervalo de Confianza al 95% sólo indica que el 95% de todos los (posiblemente infinitos) intervalos producidos por ese método frecuentista contendrán al verdadero valor (poblacional), pero ello no implica que este sea el caso de este IC95% particular, el obtenido específicamente en este experimento. En resumen: factores externos a los datos experimentales (como el tamaño muestral, que es fijado "a priori" por el investigador) son los que determinan finalmente la aceptación o no de las hipótesis, incluso en el caso de la estadística frecuentista clásica^{83,84}.

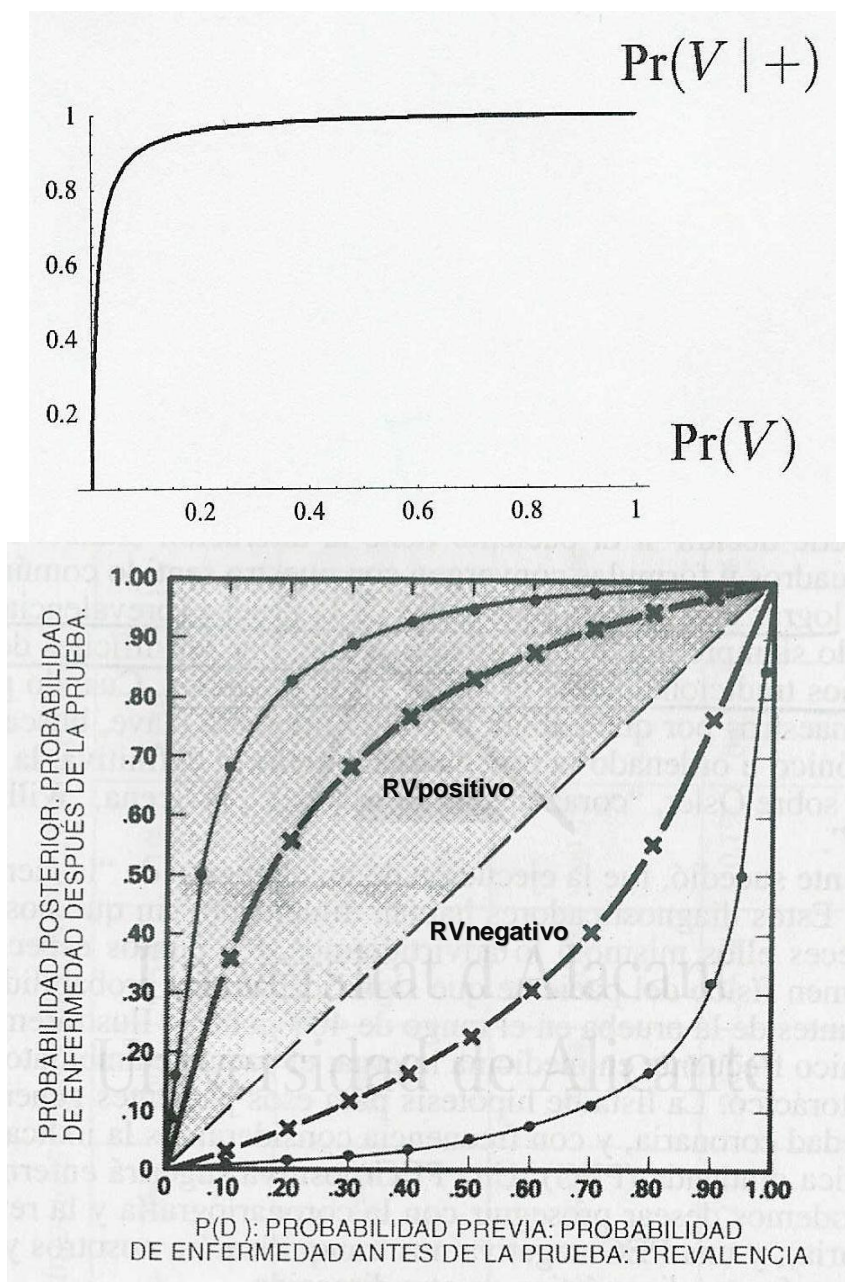


FIGURA 4: ARRIBA: Probabilidad de presentar una enfermedad cuando el test ha salido positivo (prob. a posteriori) respecto a la prevalencia de la enfermedad (prob. a priori): el gráfico demuestra la dependencia que se establece entre ambas. Cuando la creencia a priori en la enfermedad es excesivamente alta, el diagnóstico de la enfermedad es seguro, independientemente de los resultados del test. ABAJO: Zona sombreada = Resultado positivo del test; Zona blanca = Resultado negativo del test. --X-- Test con $S = 0,85$ y $E = 0,85$ ($RV+ = 5,7$ y $RV- = 0,18$). --0-- Test con $S = 0,95$ y $E = 0,95$ ($RV+ = 19$ y $RV- = 0,05$).

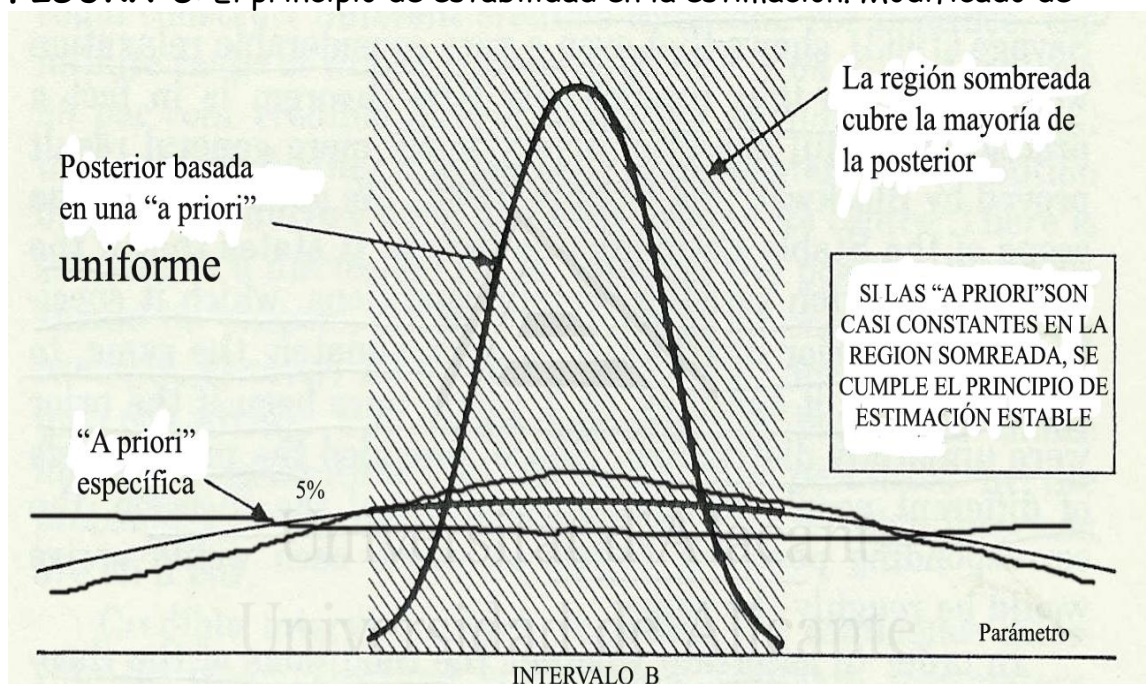
Para la mentalidad bayesiana, no sólo la objetividad pura en ciencia es inalcanzable, sino que de hecho la ciencia no aspira a ella. Al nivel básico desde el que se empieza, es obvio que la incertidumbre que sobre una hipótesis particular presenta una persona (el científico) es diferente de la que presenta otra persona: incluso en ciencia existe el desacuerdo. Es por ello importante que, incluso con su notación, la ciencia refleje esa subjetividad. Por ello se utilizan siempre las probabilidades condicionadas: si E denota el conocimiento (evidencia científica) que una persona tiene en el momento de emitir su juicio sobre una hipótesis H , mejor que $p(H)$ debería utilizarse $p(H/E)$: la probabilidad de H dada E . El científico debería usar la probabilidad como una medida de su conocimiento específico sobre un aspecto incierto, de su grado de creencia (racional) en una hipótesis. $P(H/E)$ es su creencia en H cuando conoce sólo E . Debería utilizar, al principio de su actividad científica, el concepto subjetivista o personal de la probabilidad. Luego, adquiriendo (mediante experimentación u observación) nuevos datos X , debería tender a revisar su incertidumbre: a disminuirla (si es posible) obteniendo $p(H/E,X)$ en virtud del teorema de Bayes.

Pero es que además, Edwards, Lindman y Savage han demostrado⁸⁵ el llamado **Principio de la Estabilidad en la Estimación**, que es en realidad una consecuencia práctica útil de un resultado mucho más general demostrado matemáticamente por Blackwell y Dubins⁸⁶. El teorema establece que, si las distribuciones "a priori" sobre un parámetro especificadas por personas diferentes cumplen ciertas características, cuando la cantidad/calidad de los datos experimentales incrementa progresivamente la precisión de los resultados globales de la experiencia (función de verosimilitud), las distribuciones "a posteriori" a las que llegarán todas ellas tras aprender (mediante el teorema de Bayes) de los nuevos datos serán prácticamente idénticas: finalmente estimarán el valor del parámetro con una precisión considerablemente elevada. Para que el principio de estimación estable se satisfaga, debe cumplirse que las distribuciones a priori estimadas por personas diferentes sean, aunque diferentes, relativamente "vagas e imprecisas" (no informativas) respecto de unos datos de observación muy precisos.

En ese caso, la forma y propiedades de la distribución "a priori" tiene una influencia despreciable sobre la distribución "a posteriori", que toma las características proporcionales a la función de verosimilitud.

Para asegurarnos de que las condiciones de aplicación del teorema se cumplen, debemos seguir tres pasos [FIGURA 5].

FIGURA 5: El principio de estabilidad en la estimación. Modificado de ⁶³:



Primero: debemos construir, utilizando los datos del estudio y el teorema de Bayes, una distribución "a posteriori" obtenida desde una distribución "a priori" uniforme, y comprobar que el intervalo de credibilidad al 99% o más (cuanta mayor credibilidad, mejor) bajo ella cubre un intervalo B (el delimitado por el área sombreada de la figura) de valores posibles del parámetro en cuestión. Segundo: debemos comprobar que dentro de ese intervalo B , las distribuciones "a priori" establecidas por las diferentes personas son prácticamente constantes; más específicamente, el rango de valores de probabilidad (alturas bajo cada una de las distribuciones "a priori") asignado por cada persona al intervalo B no es mayor de

un 5% (cuanto menos, mejor) del valor mínimo de probabilidad asignado dentro de B . Tercero: que el intervalo B cubre la mayor parte (por ejemplo el 95%, pero cuanto más, mejor) de la distribución "a posteriori" obtenida por cada una de las personas -combinando su opinión previa con la verosimilitud derivada de los datos-.

Esta última condición es la más difícil de comprobar en la práctica, por lo que se ha propuesto otra, de hecho más conservadora pero normalmente más fácil de verificar: que en ningún valor externo al intervalo B , las diferentes distribuciones "a priori" tomen valores astronómicamente superiores a los que toman para los valores internos a ese intervalo (se considera que un factor de hasta 1000 veces podría ser tolerable).

En otras palabras, el principio de estimación estable garantiza que, si se cumplen ciertas condiciones iniciales -circunstancias que típicamente se dan inicialmente en casi todos los casos en los que existe gran incertidumbre-, cuando la evidencia científica es abrumadora, las opiniones inicialmente discrepantes sobre un parámetro, finalmente convergen. Desde un punto de vista práctico, establece que la subjetividad de las opiniones sobre un parámetro desaparece cuando se recogen datos científicos sobre él. Dos personas con opiniones inicialmente muy discrepantes pero con mentalidades razonablemente abiertas, serán forzadas por la evidencia de una suficiente cantidad/calidad de datos científicos, a llegar a un acuerdo. Los datos fiables llegan a controlar tanto nuestras opiniones posteriores que no es necesario atender a los detalles divergentes de nuestras vagas opiniones iniciales. Ello es exactamente lo que se observa en la práctica: aunque inicialmente los científicos varíen en sus puntos de vista y discutan -incluso vehementemente- entre ellos, la evidencia abrumadora de los datos les hace finalmente acabar poniéndose de acuerdo. **La aparente objetividad de la ciencia no es otra cosa que consenso intersubjetivo.** La mentalidad bayesiana funciona exactamente como lo hace la práctica científica real.

En este sentido, para los bayesianos, la experimentación, con su capacidad de producir datos (cuya información va a codificarse como funciones de verosimilitud), es un ingrediente esencial del método científico. Éste está

constituido por una secuencia alternativa de experimentación y razonamiento, para aprender de los datos utilizando la potencia del teorema de Bayes. Bajo este punto de vista, el método científico consiste en expresar nuestras incertidumbres sobre el mundo real que nos rodea en términos de probabilidad, realizando experimentos para obtener datos, y utilizar esos datos para aprender de ellos actualizando nuestras probabilidades (creencias racionales) y con ellas nuestra visión de ese mundo incierto. El cálculo de probabilidades nos enseña el cómo. Veremos más adelante, que un impedimento grave a ese cómo se produce en situaciones en que existe carencia de buenos métodos que posibiliten la asignación de probabilidades "a priori" cuando no puede aceptarse la asunción de intercambiabilidad.

Dejemos que sea también el físico Richard Feynmann el que nos explique con su peculiar estilo, en otra parte de su famosa obra antes citada³⁷, esta visión del método científico (de nuevo la cursiva es nuestra):

"Hace algunos años, tuve una conversación con un lego sobre platillos volantes - ico como soy un científico lo conozco todo sobre platillos volantes!. Dije: 'No *creo* que existan platillos volantes'. Mi antagonista inquirió: '¿Es imposible que existan platillos volantes?, ¿Puede usted demostrar que es imposible?'. 'No', contesté. 'No puedo probar que es imposible, sólo que es muy *improbable*'. Y entonces, él dijo: '¿Y usted es científico? Si usted no puede demostrar que es imposible ¿Cómo puede decir que es muy improbable?'. Pero esa es exactamente la manera de ser del científico. Sólo es científico decir lo que es más probable y lo que es menos probable, y no estar todo el tiempo demostrando lo posible y lo imposible. Para definir lo que quiero decir, debería haberle contestado: 'Mire, quiero decir que, *desde el conocimiento que poseo del mundo que veo a mi alrededor, creo que es mucho más probable que las noticias de avistamientos de platillos volantes sean el resultado de las conocidas características irracionales de la inteligencia terrestre, que de unos esfuerzos racionales desconocidos de la inteligencia extra-terrestre*. Es sólo más probable. Eso es todo".

4.4.1.- Análisis bayesiano de Referencia:

La solución más matemáticamente rigurosa al problema de la objetividad de la ciencia, desarrollada estrictamente dentro del marco bayesiano, es el llamado *análisis bayesiano de referencia*. Dada la naturaleza del proceso bayesiano de inferencia, y tal como muestra la [FIGURA 4], en multitud de situaciones - incluso con datos experimentales extraídos de muestras grandísimas- la aplicación

del Teorema de Bayes conduce a inferencias que son completamente dependientes del estado inicial de información (prevalencia previa). Para muchos teóricos clásicos, este es el principal punto débil del bayesianismo: la dominancia de "los prejuicios y opiniones" del científico en el resultado de sus investigaciones.

Este parece, a priori, un obstáculo insalvable⁸⁷, pero la solución general no es menos ingeniosa y se debe, entre otros, al matemático valenciano José Miguel Bernardo Herranz⁵⁷: utilizar para la inferencia objetiva distribuciones "a priori" que sean de tal naturaleza que toda la información aportada al problema provenga del experimento, que minimicen la información "a priori" que entra en el proceso inferencial -los prejuicios del investigador-. Son las llamadas distribuciones "a priori" *no informativas*, representantes de la ignorancia "a priori", y que "dejan a los datos hablar por sí mismos".

Estrictamente hablando, si se utiliza el teorema de Bayes para realizar inferencias, los datos no pueden hablar completamente por sí mismos y cualquier especificación de una distribución a priori tiene alguna implicación informativa en la distribución posterior, por lo que, con rigor, no existe una distribución objetiva "a priori" que represente la total ignorancia (tal vez porque la total ignorancia no existe tampoco en los seres humanos, quienes siempre tenemos conocimientos e intuiciones). Pero los impulsores del análisis de referencia reconocen la importancia de utilizar en sus inferencias bayesianas unas distribuciones a priori que representen el concepto de que las creencias subjetivas "a priori" (prejuicios) deben tener el mínimo impacto posible -respecto al que tienen los datos muestrales- sobre el resultado final de la inferencia científica. Y de ahí la conveniencia de disponer de distribuciones "a priori" que conduzcan, vía Teorema de Bayes, a distribuciones posteriores *standard* que no incorporen las opiniones personales de los científicos. Pero además, dentro del paradigma bayesiano, las distribuciones a priori "de referencia" son compatibles con la visión subjetiva de la probabilidad ya que son sólo una herramienta técnica para obtener distribuciones posteriores "de referencia", que pueden concebirse -y de ahí su nombre- como una manera de medir la influencia de las concepciones previas sobre el resultado final.

Pueden ser vistas como una estimación de la distribución subjetiva posterior que hubiera obtenido un bayesiano que no hubiera dispuesto de (o hubiera sido incapaz de explicitar) ninguna información inicial. Cualquier bayesiano con una distribución previa subjetiva estaría interesado en comparar su propia posterior con la posterior de referencia obtenida por otro compañero desinformado.

Para obtener sus distribuciones a priori "no informativas", el análisis de referencia bayesiano utiliza toda la potencia matemática de la Teoría de la Información [**ANEXO IV**]. La complejidad técnica de su desarrollo matemático excede, con mucho, el objetivo de esta tesis, y el lector interesado puede consultarlo en el libro de Bernardo y Smith⁸⁸. En esencia, y basándose en los conceptos de "entropía mutua" y "entropía condicional" de la teoría general de la información de Shannon y de "entropía relativa" o divergencia logarítmica de Kullback y Leibler^{89,90} (llamada también información discriminante) -adaptados en 1956 por Lindley a la estadística para describir la información esperada proporcionada por un experimento⁵⁹-, se define como distribución a priori de referencia no informativa, de toda la clase de distribuciones de referencia posibles para el problema, a aquella que hace máxima la entropía (información esperada) que se pierde al pasar de la distribución a priori a la distribución a posteriori -que se gana en el sentido contrario-. Esto es, la que maximiza la divergencia logarítmica entre la distribución posterior y la previa. Para el caso particular en el que se realiza un experimento Binomial para estimar, mediante el teorema de Bayes, el valor de un parámetro θ que representa la probabilidad (subjetiva) de aparición de un evento binario, dentro de toda la familia de distribuciones a priori posibles - que corresponden a las distribuciones conjugadas con la Binomial: las distribuciones Beta- la distribución que maximiza la entropía que se gana al pasar de la distribución beta a posteriori a la distribución beta a priori es la Beta (0.5, 0.5) y no la distribución uniforme Beta (1, 1) usada por muchos bayesianos. Esta Beta (1/2, 1/2) es la que utilizaremos en esta tesis como distribución a priori de referencia (no informativa) en nuestros análisis.

4.5.-INFERENCIA BAYESIANA: LA TEORÍA MATEMÁTICA DE LA DECISIÓN:

La Teoría de la Decisión propone un determinado método matemático de tomar decisiones, porque demuestra que es *el único* racionalmente compatible con la asunción de unos pocos axiomas o principios básicos, los que caracterizan a una decisión coherente entre opciones alternativas. Así, todo su desarrollo deductivo parte inicialmente de que se asuman como ciertos "a priori" varios axiomas que describen matemáticamente lo que se entiende racionalmente como un decisor coherente. Estos axiomas son los que están enumerados en la siguiente tabla [TABLA 5], pero debemos subrayar que la adherencia a ellos garantiza la *coherencia racional* del decisor (una forma de consistencia), no la corrección de las decisiones tomadas:

TABLA 5: Los axiomas que representan el concepto de *COHERENCIA* en un decisor:

- Dados dos premios en una lotería, A y B, el decisor debe ser capaz de establecer cuál de ellos prefiere, o si es indiferente a ambos.
- Las preferencias deben ser *transitivas*: Si A es preferido a B y B es preferido a un tercer premio C, entonces A es preferido a C.
- Si A es preferido a B, y B es preferido a C, existe alguna probabilidad (p) tal que el decisor es indiferente entre la opción "obtener con seguridad B" y la opción "obtener A con probabilidad p y obtener C con probabilidad 1-p".
- Si A es preferido a B, y se da la circunstancia de que el decisor es forzado a elegir jugar en dos sorteos cuyos premios respectivos son A y B, el decisor debe preferir jugar al sorteo en el que puede obtener A con mayor probabilidad.
- Todos los sorteos en los que las probabilidades de obtener los mismos premios son iguales, son sorteos equivalentes, independientemente de si los premios se ganan como resultado de un jugada o de varias jugadas sucesivas.

Para la teoría matemática de la Decisión la única forma razonable y coherente de tomar decisiones sigue varios pasos, que están descritos en la [TABLA 6]. En primer lugar, es necesario determinar de manera exhaustiva y excluyente el número de decisiones a tomar, y, dentro de cada una, aquellos sucesos cuya ocurrencia posterior es posible y pueda modificar las consecuencias de la decisión tomada. En segundo lugar debe cuantificarse, y ya hemos visto que la estadística bayesiana demuestra que la única forma coherente de hacerlo es mediante probabilidades, la verosimilitud que el decisor otorga (subjétivamente) a la posible ocurrencia de tales sucesos. En tercer lugar, deben describirse detalladamente las posibles consecuencias de cada una de las decisiones a tomar; el decisor debe evaluar (subjétivamente) las preferencias entre esas consecuencias y cuantificarlas coherentemente en términos de una magnitud común que recibe el nombre de utilidad. Finalmente, puede demostrarse matemáticamente^{91,61} que debe tomarse aquella decisión que, en base a las probabilidades y utilidades subjétivamente asignadas, proporcione *la máxima utilidad esperada*. Cualquier desviación de estos preceptos está contradice los principios básicos de partida: cualquier otro criterio de decisión resulta inadmisibles para quien acepte tales principios de comportamiento coherente.

TABLA 6: Procedimiento para tomar decisiones racionales (coherentes):

1. Enumera las decisiones posibles: d_1, d_2, \dots, d_m .
2. Enumera los eventos posibles: e_1, e_2, \dots, e_n .
3. Asigna probabilidades a esos eventos: $p(e_1), p(e_2), \dots, p(e_n)$.
4. Asigna utilidades $u(d_i, e_j)$ a cada posible consecuencia (d_i, e_j) .
5. Elige la decisión que *maximice la utilidad esperada*

$$\text{Max } u(d_i) = \sum_{j=1}^n u(d_i, e_j) * p(e_j)$$

La Teoría de la Decisión fue desarrollada rigurosamente por Abraham Wald, uno de tantos emigrados desde Europa a los Estados Unidos después de la guerra, en su libro⁹² *Statistical Decision Functions* (1950). John Von Neumann había demostrado en 1927 el teorema del minimax para juegos finitos de suma cero, base de la teoría de juegos actual, y junto con Oskar Morgenstern⁹³, en 1944, elaboró la primera axiomática sobre la teoría matemática de la utilidad. Wald, profesor de la Universidad de Columbia, fue el primero que captó la conexión entre la teoría de juegos y la teoría estadística de comportamiento inductivo de Neyman-Pearson, y planteó el problema general de decisión como un juego con dos jugadores. Uno es del decisor, persona inteligente (coherente en el sentido de evitar el Dutch Book) que elige las estrategias del juego. El otro, representado por el azar o la naturaleza, es un jugador no inteligente pasivo, que se limita a ofrecer sus distintos estados. Al adoptar una estrategia o "decisión", se presenta un estado de la naturaleza y ello produce un resultado, una pérdida (o ganancia) que se ha de minimizar (o maximizar). Como a cada estado posible está asociada una probabilidad, las pérdidas o ganancias tendrán valores esperados. La decisión más coherente y más favorable al decisor, será la que maximice su utilidad esperada. Con ello se fundamenta el principio de la utilidad esperada, que había enunciado Daniel Bernouilli en su memoria *Specimen Theorice Novae de Mensura Sortis* que apareció en el volumen V de las Actas de la Academia de San Petersburgo, correspondiente a los años 1730 y 1731.

5.- PUNTO DE ENCUENTRO DE AMBAS EVASIONES: EL CONCEPTO DE INTERCAMBIABILIDAD Y EL TEOREMA DE REPRESENTACIÓN

Existe una propiedad compartida por ambas mentalidades estadísticas, y que las liga de un modo más íntimo de lo que toda nuestra exposición anterior podría sugerir. Se trata de la *intercambiabilidad*. A menudo, los datos disponibles

toman la forma de un conjunto $\{x_1, x_2, \dots, x_n\}$ de observaciones "homogéneas", en el sentido exacto de que sólo sus *valores* importan, y no el *orden* en el que se han practicado las mediciones. Esta idea intuitiva es la captada formalmente por el concepto de intercambiabilidad: una secuencia o conjunto de variables (vectores) aleatorios (x_1, x_2, \dots, x_n) es, **para ti**, intercambiable bajo las **condiciones K**, si tu distribución conjunta de ellos dadas K, es invariante bajo permutación de los sufijos. Por ejemplo, $p(x_1=3, x_2=5 / K) = p(x_2=3, x_1=5 / K)$ tras permutar 1 y 2. Una secuencia infinita es intercambiabile si cada una de las posibles subsecuencias finitas de ellas lo son. Se ha subrayado el "para ti" y "en las condiciones K" porque la intercambiabilidad de las variables **es un juicio subjetivo**, y tú puedes cambiar de opinión sobre la intercambiabilidad si las condiciones cambian.

En 1930⁹⁴ Bruno de Finetti publicó un resultado sensacional: el llamado **teorema de representación**. Demostró que si **juzgamos subjetivamente** un conjunto (x_1, x_2, \dots, x_n) de variables de Bernoulli -variables dicotómicas que sólo toman el valor 0 o 1- como intercambiables, se cumple que su distribución conjunta tiene un representación integral dependiente del parámetro θ de la forma

$$p(x_1, x_2, \dots, x_n / K) = \int \prod_{i=1}^n p(x_i / \theta) p(\theta / K) d\theta = \int \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1 - x_i} p(\theta / K) d\theta$$

$$\text{siendo } \theta = \lim_{n \rightarrow \infty} r/n \text{ ,}$$

donde $r = \sum x_i$ es el número de pruebas positivas. Los resultados de De Finetti se han extendido recientemente a situaciones mucho más generales^{95,96}. Así, para sucesiones de variables aleatorias arbitrarias $(y_1, y_2, \dots, y_n / K)$, la intercambiabilidad en las condiciones K conlleva representaciones integrales de la forma

$$p(y_1, y_2, \dots, y_n / K) = \int \prod_{i=1}^n p(y_i / \psi) p(\psi / K) d\psi$$

$$\text{siendo } \psi = \lim_{n \rightarrow \infty} f(y_1, y_2, \dots, y_n) \text{ .}$$

Aquí $p(y_i / \psi)$ denota *algún* modelo probabilístico de los datos muestrales, el parámetro ψ es el límite (cuando $n \rightarrow \infty$) de alguna función de los datos observados $f(y_1, y_2, \dots, y_n)$, y $p(\psi / K)$ es alguna distribución "a priori" de probabilidad sobre el parámetro ψ en las condiciones K.

Nótese que el teorema de representación general es un **teorema de existencia**: no especifica un modelo probabilística determinado ni una distribución a priori sobre el parámetro. Las asunciones adicionales necesarias para especificar un modelo muestral particular se describen en cada teorema de representación particular. Así, en el caso particular del modelo binomial (variables dicotómicas de Bernoulli) -que fue el demostrado originalmente por de Finetti-, el parámetro requerido es θ , la proporción (asintótica) de pruebas positivas. En cada caso particular se requiere un esfuerzo adicional para decidir la distribución "a priori" específica mas adecuada.

En esencia, el teorema de representación general -un resultado puro de la teoría matemática de probabilidades- demuestra que, si tú juzgas que una secuencia de variables aleatorias es (infinitamente) intercambiable, **debes** considerar que para tí existe un modelo o estructura probabilística subyacente al problema que describe el mecanismo aleatorio que ha generado los datos. Este modelo hace que ese conjunto de observaciones constituyan una **muestra aleatoria**, y que los miembros de la secuencia (los datos obtenidos x_i) puedan considerarse entre sí independientes e idénticamente distribuidos (i.i.d.) según ese modelo -en este sentido, intercambiabilidad implica independencia. El modelo probabilístico subyacente al problema estará caracterizado por un parámetro ψ , y el teorema también demuestra que ese parámetro ψ está definido como el límite (cuando $n \rightarrow \infty$) de alguna función de los datos observados $f(y_1, y_2, \dots, y_n)$. Es decir, que el parámetro ψ , que ordinariamente está constituido por elementos (θ, α) de los que sólo nos interesará alguno para solucionar nuestro problema particular, está determinado por **las propiedades frecuentistas de la secuencia**. Así, para el caso simple de una secuencia de variables de Bernoulli, ψ es la **proporción** límite (cuando $n \rightarrow \infty$) de ellas que toman valor 1: el límite de las frecuencias relativas r/n . Por tanto, asumiendo intercambiabilidad, el parámetro ψ tiene un valor (frecuentista) fijo, objetivo o "empíricamente verdadero", aunque sólo es comprobable asintóticamente.

Además el teorema de representación general también demuestra que, si tú asumes intercambiabilidad en la secuencia, para ti *debe existir* una distribución de probabilidad "a priori" $p(\psi / k)$ sobre el parámetro -en las condiciones K- que caracteriza al modelo probabilística y describe el mecanismo aleatorio que ha generado los datos. Esto es, la información disponible, en las condiciones K, relativa a la incertidumbre (grado racional de creencia) sobre los valores del parámetro ψ , debe describirse necesariamente (como ya hemos visto que demuestra también el Teorema Ramsey-de Finetti) por *alguna* distribución de probabilidad $p(\psi / k)$. Lo importante es que asumiendo que la secuencia tiene la simple propiedad de intercambiabilidad (el orden no importa) -de hecho un asunción mucho más débil que la que hace en este caso la estadística frecuentista, la asunción de un proceso de Bernoulli- se produce la distribución "a priori" $p(\psi / k)$ y se garantiza que se pueda realizar un análisis bayesiano del problema⁹⁷. No existe pues, tal y como muchos críticos de este tipo de enfoque aseguran, la asunción de una distribución *a priori*.

Como hemos visto anteriormente (pág. 20), el concepto de muestra aleatoria de observaciones i.i.d. es el concepto central de la experimentación en la teoría estadística clásica: en base a la ley de grandes números, la aleatorización de muestras grandes permite minimizar la confusión y justifica la inferencia causal. Incluso cuando es obviamente inapropiado, como en el estudio de las series temporales, la mayoría de textos de estadística y los modelos clásicos de diseño de experimentos sólo se ocupan de datos extraídos aleatoriamente tras repeticiones del fenómeno y usan las observaciones i.i.d. como base de todos los cálculos. Con la asunción de i.i.d. las ideas frecuentistas son adecuadas, pero eso ha hecho desarrollar una creencia de que la probabilidad debe estar siempre basada en la frecuencia. Ello restringe necesariamente el rango de actividades experimentales en las cuales puede hacerse un uso apropiado de las técnicas estadísticas clásicas: aquellas en las que sea posible la repetición del fenómeno y el muestreo aleatorio. De ahí los graves problemas derivados de realizar inferencia causal con muestras no aleatorizadas o datos retrospectivos (confusión, interacción, etc...), que han

dado pié a la utilización, por ejemplo en la epidemiología moderna, de los modernos métodos multivariantes^{98,99,100,101} buscando minimizar al máximo esos sesgos, uso por otra parte no exento de riesgos¹⁰². La confusión entre frecuencia y probabilidad niega a los frequentistas la posibilidad de usar la probabilidad como una medida de incertidumbre, con el resultado de que ha sido necesario desarrollar para ello conceptos matemáticamente incoherentes como los intervalos de confianza.

Por ello, el teorema de representación es uno de los resultados más extraordinariamente sobresalientes de la teoría de la probabilidad y del pensamiento estadístico moderno, pues constituye el punto de encuentro de las mentalidades frequentista y bayesiana. Es muy general, y sin embargo muy simple. Pero es muy importante por su enorme aplicabilidad: las secuencias intercambiables son tremendamente frecuentes en la práctica⁹⁷. Si un bayesiano juzga que una secuencia de variables aleatorias es intercambiable entre sí (su orden no importa), está efectivamente haciendo el mismo juicio sobre los datos que hace un frequentista cuando considera que esta secuencia es una muestra aleatoria -esto es, obtenida mediante un mecanismo de aleatorización- de un fenómeno repetido, pero con la adición de una (necesaria) especificación de una distribución de probabilidad "a priori" sobre el parámetro. Esto tiene dos importantes consecuencias positivas.

En primer lugar, ello amplía el horizonte de actuación de la estadística a campos en los que aparece incertidumbre, pero en los que la repetición de los eventos es imposible. Es el caso en el que nos enfrentamos a acontecimientos simples o casos únicos, y no tiene sentido hablar de repeticiones: las probabilidades no pueden ser entendidas como frecuencias. Es lo que ocurre, por ejemplo, en el terreno de la aplicación de la justicia o la ciencia forense. Durante los juicios, existe incertidumbre sobre la culpabilidad del acusado, incertidumbre que se va atemperando mediante las pruebas aportadas con la esperanza de alcanzar finalmente un consenso sobre la culpabilidad o no del acusado. Esto, como vemos, coincide con el proceso de inferencia que constituye el teorema de Bayes. Para aplicar justicia no nos sirve un método de detectar culpabilidad que acierte un

95% de las veces que se aplica en las mismas condiciones (intervalo de confianza): necesitamos un método que nos ayude en este caso determinado, con este sospechoso determinado, y con las circunstancias especiales -e irrepetibles- del día de autos. La estadística bayesiana, en base al principio de intercambiabilidad, da respuestas apropiadas a las situaciones únicas. En este caso particular, sin necesidad de repeticiones, nos da la probabilidad de que *ese* sospechoso sea culpable de *ese* crimen. No necesitamos incluir este caso en una secuencia de casos similares para obtener un veredicto que se justifique "asintóticamente".

La segunda consecuencia no es menos interesante. La intercambiabilidad permite al bayesiano utilizar lo que Ian Hacking llama el *principio de frecuencia como creencia*¹⁹, que conecta las probabilidades tipo frecuencia con las de tipo creencia. Inicialmente, este principio es una regla de actuación en condiciones de ignorancia. Supongamos que conocemos la frecuencia de aparición de un resultado en experimentos de n repeticiones del mismo evento, y sin embargo somos completamente ignorantes sobre el resultado o las condiciones en las que se va a producir una nueva repetición aislada del mismo fenómeno. Si no tenemos ninguna información previa, podemos juzgar subjetivamente los eventos pasados y futuros como intercambiables entre sí: una intercambiabilidad subjetiva basada sobre una asunción de ignorancia. En ese caso, nuestra creencia o incertidumbre *a priori* sobre el resultado final de una única repetición futura del evento puede ser medida por esa frecuencia de aparición en el pasado, que ya conocemos. Cuando no tengamos más información, podemos actuar como bayesianos utilizando las probabilidades frequentistas (obtenidas en repeticiones pasadas del mismo evento) como grados de creencia para medir nuestra incertidumbre *a priori* en el resultado futuro del evento. Es interesante subrayar cómo esta asunción de intercambiabilidad de los elementos del presente con el pasado ya la habíamos visto en boca de Hume como imprescindible para posibilitar cualquier inducción (pág. 19).

La mayoría de las investigaciones científicas son observacionales y no experimentales. El uso de técnicas de estadística frequentista en tales escenarios

externos a la intercambiabilidad puede traernos muchos problemas, y lo normal es que aparezcan factores que produzcan confusión a la hora de realizar inferencias causales. Por ello, siempre -incluso usando métodos frecuentistas- es esencial exigir un cierto juicio de intercambiabilidad, y para ello se realiza el ajuste multivariable de los resultados. Los juicios de intercambiabilidad en contextos frecuentistas se expresan como juicios sobre "subpoblaciones" a las que pertenecen las unidades de que estamos tratando: las probabilidades condicionadas o ajustadas. Una vez se ha realizado el ajuste adecuado, el principio de *frecuencia como creencia* posibilita también que las frecuencias ajustadas puedan utilizarse como medidas de incertidumbre o grados de creencia bayesianos para aplicarlas sobre el resultado de un único evento simple considerado subjetivamente intercambiable con los otros. En ese sentido, la intercambiabilidad es una poderosa extensión del concepto tradicional de muestra aleatoria. Lo importante es reconocer que la intercambiabilidad es *siempre* un juicio que debe realizar el investigador, no una propiedad del mundo externo, en el mismo sentido en que la causalidad es un juicio nuestro sobre el mundo, no una verdad inherente a él¹⁰³.

Por otra parte, tal y como discute cuidadosamente Rubin¹⁰⁴, el argumento bayesiano se simplifica considerablemente si la intercambiabilidad no sólo se juzga subjetivamente, sino se asegura "intersubjetivamente" utilizando un mecanismo aleatorio de muestreo. La aleatorización no es necesaria, pero sí muy útil en inferencia bayesiana: un mecanismo de aleatorización hará que mucha más gente, no sólo el investigador, juzgue a la secuencia como intercambiable, y por tanto i.i.d. La posibilidad de una intercambiabilidad más objetiva es lo que dá la superioridad a los diseños experimentales, incluso cuando se aplican técnicas bayesianas.

6.- LOS ÍNDICES DE PREDICCIÓN CLÍNICA COMO SUSTENTO DE LA TOMA DE DECISIONES EN MEDICINA

Determinar el diagnóstico y establecer el pronóstico de un paciente son dos actividades íntimamente relacionadas que constituyen el centro de la práctica de cualquier médico. Condicionan seriamente las decisiones que éste toma sobre las recomendaciones que realiza a las personas que trata: las pruebas que va a ordenar, el pronóstico que va a predecir, el tratamiento que se debe aplicar... Es claro que el principal papel del médico es tomar decisiones¹⁰⁵. Pero ya hemos visto como todas las decisiones médicas -como las de cualquier científico- se toman en un ambiente de incertidumbre e inseguridad. Cada vez más, los médicos toman conciencia de importancia de la situación y, aunque en su educación clásica rara vez se les instruye, están aprendiendo a aplicar en su quehacer diario técnicas de decisión basadas en el enfoque formal de la teoría matemática de la decisión, adaptado de los mundos militar y mercantil¹⁰⁵.

Nuestra experiencia en la práctica clínica nos dota de una sensación intuitiva (el juicio u "ojo" clínico) de cuales hallazgos de la anamnesis, la exploración física o los exámenes complementarios son cruciales para hacer un diagnóstico certero y predecir adecuadamente un pronóstico. Pero, desgraciadamente, la experiencia también nos enseña que esta misma intuición resulta a veces muy engañosa: como humanos estamos inclinados a dirigir nuestra atención hacia lo nuevo, lo raro, lo interesante o lo emocionalmente atractivo y no somos muy buenos haciendo observaciones sistemáticas, insesgadas y consistentes en el tiempo¹⁰⁶.

Así, todos conocemos cómo una experiencia "mala" con un paciente individual puede condicionar negativamente la práctica futura de un médico, incluso si esa experiencia fue una excepción rara. O cómo, por el contrario, si un enfermo responde muy bien a una medicación, somos propensos a creer que también funcionará con el próximo paciente. Nuestra vivencia durante el periodo de

formación, a menudo en grandes hospitales terciarios, pudo predisponernos a convivir con la patología rara o muy abigarrada, y ello suele traer como resultado que sobreestimemos la probabilidad de encontrarnos de nuevo esas formas de enfermedad.

La disponibilidad de información, como la lectura reciente de un artículo en una revista médica, aumenta la conciencia que tenemos de un determinado proceso o condición y hace también más probable que (correcta o incorrectamente) lo diagnostiquemos en los próximos pacientes. Además, la opinión de nuestros colegas o la literatura médica no aprehendida críticamente puede estar más basada en anécdotas que en datos objetivos sobre prevalencia y/o evolución de una enfermedad. Estos y otros sesgos, que no son sino la traslación a la medicina del "problema de la inducción" en ciencia, nos privan de capacidad para acertar siempre con nuestras predicciones clínicas.

Para ayudar a establecer objetivamente un pronóstico, basándose en la experiencia acumulada durante años por clínicos de gran prestigio, los médicos de la antigüedad desarrollaron *reglas de predicción* que expresaban en forma de aforismo¹⁰⁷. Así, por ejemplo, los médicos del Egipto arcaico (3er. Milenio a.d.C.) dejaron escrito:

"Si examinas a un hombre que tiene una rotura en la cámara de la nariz y encuentras su nariz torcida, su cara desfigurada y una hinchazón por encima que sobresale, dirás acerca de él: una enfermedad que voy a tratar.

Si examinas un hombre que tiene una herida abierta en la cabeza, que penetra en el hueso, fractura el cráneo y deja el cerebro al descubierto, deberás palpar su herida. Comprobarás si la fractura que tiene en su cráneo es semejante a los pliegues que se forman en el cobre fundido, y si palpita y cede bajo tus dedos como la parte débil de la coronilla de un niño antes de que se suelde. Cuando suceda que no palpite ni ceda bajo tus dedos mientras que el cerebro está al descubierto y el paciente arroja sangre por ambas fosas nasales y tiene rigidez en el cuello, dirás acerca de él: una enfermedad que no es posible tratar¹⁰⁸".

De una época más cercana a la nuestra (siglos V-IV a.d.C.) data la descripción clásica de la "facies hipocrática"¹⁰⁹:

"Conviene investigar así en las enfermedades agudas. Primeramente observar la cara del enfermo, si es semejante a la de los sanos, sobretudo a la del mismo enfermo cuando tenía salud; porque esto sería lo mejor, y cuanto más diste de lo semejante tanto será más temible. Por ejemplo: la nariz afilada, hundidos los ojos, caídas las sienes, frías y encogidas las orejas y sus pulpejos retorcidos, y dura la

cutis del rostro y tirante y árida, y la color de todo el semblante amarilla y amoratada. Si tal se presenta el semblante al comienzo de la enfermedad, sin que todavía por las demás señales puedan hacerse conjeturas, conviene preguntar - desde luego- si el enfermo estuvo desvelado, si padeció abundantes cámaras o si tiene por ventura mucha hambre. Cuando confesare alguna de estas cosas debe tenerse por menos de cuidado, juzgándose de todos modos en un día y una noche si por aquellas causas tiene tal apariencia el semblante. Pero si nada de esto confiesa ni en el tiempo dicho se compone su rostro, entiéndase que es señal de muerte segura".

Tales muestras de experiencia destilada son memorables. Sin embargo, simplifican en exceso la compleja realidad de los problemas individuales de la práctica clínica diaria, y evitan que valoremos adecuadamente toda la riqueza de detalles o matizaciones con que se nos plantean las situaciones reales¹⁰⁷. Ello les resta eficacia como herramientas que guíen el proceso de toma de decisiones en condiciones de incertidumbre que constituye el meollo de nuestra actividad médica cotidiana.

En la medicina científica actual¹¹⁰, se llama **regla o índice de predicción clínica** (en inglés *clinical prediction rule* o *predictive index*) a una herramienta matemática de uso clínico capaz de ponderar cuantitativamente la contribución individual que varios componentes de la anamnesis, la exploración física y los resultados de análisis básicos de laboratorio tienen sobre el diagnóstico, el pronóstico y/o la respuesta probable al tratamiento de un paciente individual. En general, las modernas reglas o índices de predicción clínica (en adelante IPC) han sido desarrolladas en estudios realizados con bases de datos derivados de miles de pacientes y utilizando métodos matemáticas de análisis multivariable tremendamente sofisticados¹¹¹. En la literatura internacional también se conocen como *clinical prediction guides*, o *clinical decision rules*. "Predicción" indica su poder de ayudar al médico a avanzar o adivinar con antelación la aparición de un evento clínico futuro; "Decisión" implica la capacidad para ayudar al médico a elegir entre una o varias alternativas de acción diagnóstica o terapéutica.

Los estudios sobre una IPC son estudios para establecer un pronóstico, pero desde el punto de vista epidemiológico son estudios superponibles a los que se realizan sobre métodos diagnósticos. La naturaleza de la relación entre la variable

predictora (el test diagnóstico o el factor pronóstico) y la de resultado (presencia o ausencia de enfermedad, aparición o no del resultado) raramente, si acaso lo es alguna vez, es causal. El resultado del test o la presencia del factor pronóstico suele ser la *consecuencia* del proceso patológico, no su causa, o una situación intermedia en la cadena causal entre la enfermedad y su resultado p. ej la muerte. Por eso, en la mayoría de situaciones clínicas reales, el límite entre diagnóstico y pronóstico se borra completamente hasta desaparecer. La aplicación de un IPC a veces condiciona una decisión terapéutica (p. ej. decidir a que pacientes con lesiones traumáticas menores de tobillo debe practicarse una radiografía^{112,113,114}), otras veces establece una predicción sobre un pronóstico (p. ej. cuál es la probabilidad de que un paciente con sospecha de enfermedad coronaria muera durante los próximos 4 años¹¹⁵) y otras muchas veces provee al médico de una "Probabilidad post-prueba" o de una "Razón de verosimilitudes" para su aplicación en un problema de diagnóstico diferencial. Aunque el nombre correcto de estos últimos debería ser *índice, regla o guía de diagnóstico clínico*, usaremos siempre las siglas IPC independientemente de que su resultado nos sugiera una estrategia diagnóstico-terapéutica apropiada, nos dé la probabilidad estimada de un evento clínico futuro o nos indique una variación en la verosimilitud de un determinado diagnóstico.

Cualquiera que sea el resultado generado, los actuales IPCs demuestran su verdadero potencial cuando se utilizan en situaciones clínicas en las que el proceso de toma de decisiones es muy complejo, los riesgos clínicos potenciales son muy altos, o existe oportunidad de aumentar la eficiencia, ahorrando costes sin comprometer la eficacia del cuidado de los pacientes. Pero para ello es muy importante que estén bien desarrollados, validados, y se haya comprobado su impacto real en la excelencia de la atención sanitaria.

En el [**ANEXO VII**] se detalla con un ejemplo cómo se usa correctamente una prueba diagnóstica, desde el punto de vista clínico. Se trata del modelo de Pauker-Kassirer, y en un trabajo nuestro¹¹⁶ se utiliza este modelo para

ilustrar la utilización de una IPC incorporándola adecuadamente al proceso de toma de decisiones "a pie de cama".

7.- USO DE LA EVASIÓN BAYESIANA EN MEDICINA CLÍNICA: JUSTIFICACIÓN DE ESTA TESIS DOCTORAL

Cuando se enfrenta a los problemas de diagnóstico, la Medicina Clínica utiliza la evasión bayesiana del problema de la inducción: el médico aplica el teorema de Bayes partiendo de la concepción de la probabilidad como una medida del grado de creencia (racional) sobre la presencia, en el paciente que tiene delante, de una enfermedad.

Cuando tiene que realizar un diagnóstico, el médico se enfrenta a una situación en la que debe tomar decisiones coherentes en ambiente de incertidumbre, pero en la que la repetición de los eventos muestrales es imposible. Debe dar un diagnóstico ante un paciente concreto y en un momento determinado. La actividad diagnóstica trata, pues, de acontecimientos singulares y casos únicos - el enfermo particular que tiene delante en las circunstancias propias e irrepetibles de ese día- y no tiene sentido utilizar un método que necesite incluir este caso en una serie de casos similares para obtener un veredicto que se justifique "asintóticamente" (que funcione el 95% de las veces que se use)¹¹⁷. La estadística bayesiana, en base sólo al principio de intercambiabilidad, da respuestas apropiadas a las situaciones únicas; da al médico la probabilidad (grado de creencia racional) de que *ese* enfermo concreto sea portador, en ese momento, de *esa* enfermedad. Así pues, en el proceso de diagnóstico, para la Medicina Clínica resulta indispensable^{118,119} la estadística bayesiana: la forma racional de tomar decisiones científicas de una manera coherente en situaciones de incertidumbre, sin que se requiera de la repetición de una serie de eventos muestrales ni del muestreo aleatorio.

Es por ello sabido que, para evaluar el rendimiento clínico de una prueba diagnóstica, la epidemiología utiliza unos índices (Sensibilidad, Especificidad, Razones de Verosimilitud, Área bajo la Curva ROC) cuya interpretación sólo tiene sentido bajo la perspectiva bayesiana¹²⁰. Y por eso llama poderosamente la atención que se utilicen rutinariamente métodos clásicos de inferencia frecuentista (intervalos de confianza, pruebas de significación estadística) -es decir, basados en distribuciones muestrales que asumen aleatorización de una serie grande de eventos repetidos- para evaluar la potencia de dichos índices o para comparar entre sí dos métodos de predicción. Así, por ejemplo, cuando se quiere evaluar el rendimiento diagnóstico de un modelo de regresión logística como predictor de mortalidad, el nivel de significación estadística de la Sensibilidad, la Especificidad¹²¹ y de las Razones de Verosimilitud¹²² se evalúa utilizando el teorema del Límite Central (la aproximación normal a la distribución binomial de la proporción muestral), la significación estadística del Área bajo la curva ROC¹²³ se obtiene usando el Test no paramétrico de Wilcoxon (o la versión de Mann-Whitney de este)¹²⁴, y la calibración del modelo se mide con el test de Hosmer-Lemeshow cuya significación estadística se obtiene con la distribución muestral de Ji-Cuadrado¹²⁵. Si lo que se quiere es comparar entre sí dos modelos de regresión, se utiliza la significación estadística de la diferencia entre sus Áreas bajo la curva ROC, que se obtiene habitualmente con el método de Hanley-McNeil¹²⁶ basado en la distribución muestral normal Standard. Sin embargo, y aquí radica a nuestro juicio el error, no se exige como imprescindible que los datos con los que se van a extraer esos índices de evaluación de las pruebas diagnósticas se hayan obtenido mediante un muestreo aleatorio. A nuestro entender, ello *matemáticamente es un sin sentido*. El presente trabajo se justifica para intentar resolver esta paradoja.



III. - Objetivos.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

Para estimar la precisión estadística de los índices de rendimiento clínico que evalúan la exactitud de un test diagnóstico se utilizan habitualmente los métodos estadísticos clásicos: nivel de significación e Intervalos de Confianza. La validez de estos métodos frecuentistas se establece en base al Teorema del Límite Central, por lo que requiere necesariamente de un muestreo aleatorio y de una justificación asintótica. El objetivo principal de este trabajo es proponer como alternativa la aplicación de métodos estadísticos sencillos pero estrictamente bayesianos. Bajo el supuesto de intercambiabilidad, la validez de los métodos estadísticos bayesianos no requiere de la repetición de una serie de eventos muestrales ni del muestreo aleatorio.

Este objetivo pretende alcanzarse a través de la aplicación práctica a la comparación de la exactitud diagnóstica de dos test de predicción de mortalidad, un modelo de regresión logística y una red neuronal artificial, aplicados sobre los datos obtenidos con el "índice de predicción clínico" más utilizada en nuestro medio para estimar la mortalidad en la UCI de pediatría. El objetivo secundario es estimar si la exactitud diagnóstica es superior en alguno de los dos modelos, tanto desde el punto de vista de su capacidad discriminante como de su calibración.



Universitat d'Alacant
Universidad de Alicante



IV. - Hipótesis.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

La hipótesis principal es que, bajo el supuesto de intercambiabilidad y sin que se requiera de un muestreo aleatorio ni una justificación asintótica, es posible estimar la exactitud de un test diagnóstico mediante índices de rendimiento cuya precisión se establezca con métodos estadísticos sencillos pero estrictamente bayesianos.

La hipótesis secundaria es que la exactitud diagnóstica del test basado en un modelo de Red Neuronal Artificial es superior a la del modelo de Regresión Logística.



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante



V. - Pacientes, material y métodos.



Universitat d'Alacant
Universidad de Alicante

1.- LA PUNTUACIÓN EN EL ÍNDICE "PRISM"

El score PRISM (Pediatric Risk of Mortality)¹²⁷ es una regla de predicción clínica que se usa habitualmente en Cuidados Intensivos Pediátricos para predecir la mortalidad de los pacientes ingresados en UCIP utilizando los valores de 14 variables fisiológicas recogidas durante las primeras 24 horas del ingreso del niño [ANEXO V]. El rango total de la puntuación es de 0 a 76 puntos. En el diseño original del test, la mortalidad del niño se predice ajustando un modelo de regresión logística que usa como variables predictoras la edad (meses), la peor puntuación en el score PRISM de las primeras 24 horas y la presencia o ausencia de cirugía previa al ingreso (1 postoperados, 0 no postoperados). Las condiciones de aplicación y las características de medición de las distintas variables que forman el test están bien descritas en los textos de uso frecuente en la literatura de los cuidados intensivos pediátricos¹²⁸.

El test ha sido desarrollado en Estados Unidos con los datos independientes de cuatro UCIP de hospitales terciarios¹²⁷, pero validado prospectivamente en una cohorte distinta de niños de una UCIP Escocesa¹²⁹. El rendimiento diagnóstico fue similar al obtenido en el estudio americano, con una sensibilidad mejor en los pacientes no postoperatorios (71%) que en los postoperatorios (17%), pero con una especificidad excelente en ambos grupos (96% vs. 100%)¹³⁰. El score se usa habitualmente en la actividad asistencial diaria de las UCIPs españolas, pues el grupo de la UCIP del Hospital Central de Asturias ha publicado una validación con una cohorte prospectiva de niños de nuestro país^{131,132}.

2.- DISEÑO DEL ESTUDIO

Se trata de un estudio observacional analítico de cohortes retrospectivas para obtener y comparar dos métodos diagnósticos de predicción de mortalidad tras el ingreso en UCIP. Uno es un modelo clásico de regresión logística y el otro es una red neuronal artificial, y ambos están basados en *tres variables predictoras*: la edad en meses, la puntuación en el índice PRISM, y la presencia o ausencia de cirugía previa al ingreso (1 postoperados, 0 no postoperados).

El estudio se desarrolla en tres fases: una primera de desarrollo de los tests, una segunda de validación, y finalmente una tercera de comparación del rendimiento diagnóstico de ambos tests entre sí. Durante cada una de ellas, la capacidad de discriminación y la calibración de cada test será medida utilizando las herramientas estadísticas frecuentistas clásicas, y se va a proponer la utilización de una evaluación realizada mediante sus alternativas bayesianas equivalentes.

3.- 1ª FASE: DESARROLLO DE LOS TESTS

3.1- MODELO PREDICTIVO DE REGRESIÓN LOGÍSTICA:

3.1.1.- Diseño:

Estudio observacional analítico de una cohorte retrospectiva.

3.1.2.- Muestreo y tamaño muestral:

El estudio se realiza sobre el análisis retrospectivo de la base de datos de todos los pacientes ingresados en la UCIP del Hospital Central de Asturias desde su inauguración en Octubre de 1995 hasta el mes de Diciembre de 2003. La descripción de la actividad asistencial y de las características de la unidad está recogida y bien documentada en la literatura especializada de nuestro país, en la que se ha publicado también una validación prospectiva del score PRISM^{133,134}.

No está bien establecido cuál es el tamaño muestral adecuado para ajustar un modelo de regresión logística, pero una regla empírica muy extendida para obtener estimaciones estables de los coeficientes de regresión, es utilizar 10 pacientes por cada evento (muerte) producido en la variable dependiente a predecir¹³⁵. La regla se basa en los estudios de simulación de Peduzzi y Feinstein¹³⁶, y es la que hemos seguido para calcular el tamaño muestral de este estudio: la UCIP del Hospital Central de Asturias tiene unos 250 ingresos anuales¹³⁷ y tiene una mortalidad del 3% de estos pacientes¹³². La muestra total de nuestro estudio incluyó 43 exitus, por lo que el tamaño muestral mínimo adecuado se estimó en 500 pacientes.

3.1.3.- Ajuste del modelo predictivo de Regresión Logística:

El modelo de regresión logística (RL) se ha ajustado utilizando el programa estadístico SPSS 11.0.1. mediante el procedimiento clásico habitual, que describiremos brevemente. Los parámetros de cada variable del modelo se han determinado mediante estimación de máxima verosimilitud, usando el algoritmo de Newton-Raphson. No se han usado los límites de convergencia que emplea el programa por defecto, sino que estos se han fijado, tanto el inferior como el superior, en 0'0001 y se ha realizado un máximo de 50 iteraciones.

Se parte inicialmente de las tres variables predictoras comentadas anteriormente: la edad en meses, la puntuación en el índice PRISM, y la presencia o ausencia de cirugía previa al ingreso, codificada binariamente (1 postoperados, 0 no postoperados). La selección final de las variables que forman el modelo final se ha realizado utilizando para ello el procedimiento de introducción secuencial basado en el cociente de verosimilitud (forward selection by RV) y el resultado se ha comprobado de forma independiente repitiendo el proceso mediante el procedimiento de retirada secuencial según el cociente de verosimilitud (backward selection by RV) con probabilidades de entrada de 0'05 y de retirada de 0'1. Finalmente, se ha seleccionado el modelo más parsimonioso formado sólo con las variables cuyos coeficientes mostraron significación estadística, cuyo nivel se fijó previamente en el habitual del 5%.

3.1.4.- Evaluación de la exactitud diagnóstica del modelo:

La calidad del modelo como método de predicción se ha evaluado midiendo su poder de discriminación y su calibración. En esta fase sólo realizaremos una evaluación de la calidad del modelo como método predictivo con la muestra utilizada para su ajuste (validación interna). Es la llamada por algunos teóricos la capacidad "post-dictiva" del modelo¹³⁸.

3.1.4.1.- Capacidad discriminante:

La capacidad de discriminación es una medida de lo bien que el modelo clasifica a los individuos en las dos categorías: muertos y vivos. Las medidas clásicas comúnmente usadas para evaluar la capacidad discriminante son la Sensibilidad, la Especificidad, los Valores Predictivos y las Razones de Verosimilitud para positivos y para negativos, la Odds Ratio Diagnóstica y el área bajo la curva ROC (AUC). Para calcular estos índices, se ha determinado el punto óptimo de corte como aquel que corresponde al punto más cercano de la curva ROC al extremo superior izquierdo del gráfico considerando, por simplicidad, que tiene el mismo impacto un falso positivo que un falso negativo. Con una hoja de cálculo convencional diseñada en MS-Excel se han estimado los Intervalos de Confianza al 95% de estos índices. La significación estadística de la diferencia entre el AUC y la predicción esperada por azar (correspondiente a una AUC de 0.5)¹³⁹ se obtiene con el programa EpiDat 3.1 mediante la versión de Mann-Whitney del Test no paramétrico de Wilcoxon¹⁴⁰. Para el cálculo del Intervalo de Confianza al 95% del AUC, se ha usado el método de Hanley y McNeil^{140,126} -que asume normalidad-, y el método no paramétrico de DeLong¹⁴¹, usando el programa EpiDat 3.1.

Para evaluar desde el punto de vista **bayesiano** la capacidad discriminante del modelo, se ha utilizado el teorema de Bayes y se ha realizado un análisis de referencia para estimar los Intervalos de Credibilidad al 95% de la Sensibilidad, la Especificidad, los Valores Predictivos y el AUC. Es bien sabido que Sensibilidad Especificidad y Valores Predictivos son probabilidades, y que el área bajo la curva ROC (AUC) se define como *la probabilidad* de que el modelo clasifique correctamente un par de individuos vivo y muerto, pues asigne mayor probabilidad de muerte al individuo que finalmente morirá^{123,124}. Nuestra propuesta consiste en obtener, mediante el teorema de Bayes, la distribución a posteriori sobre la Sensibilidad, la Especificidad y el AUC, partiendo de la distribución a priori menos informativa para el problema: la Beta (0.5, 0.5)^{57,142}. Los datos muestrales se introducen codificados mediante la función de verosimilitud de un experimento binomial, la verosimilitud conjugada con la distribución Beta^{143,144,145}, y la

probabilidad posterior se expresa con la distribución Beta. Bajo ella, se calculan los intervalos de Credibilidad al 95% del verdadero valor de cada uno de los índices.

Cuando la distribución posterior de probabilidad sobre un parámetro es una función muy asimétrica, un intervalo a ambos lados de la media que deje por fuera áreas iguales de ambas colas contendrá generalmente algunos valores del parámetro en el que estemos interesados que tendrán, de hecho, menos probabilidad posterior que algunos valores por fuera de dicho intervalo. Este problema se evita usando los llamados Intervalos o regiones de Máxima Densidad (IMD) de probabilidad (Highest Probability Density, HDP). Se calculan de tal manera que son idénticas las ordenadas de los valores límite del intervalo, y por ello son los intervalos más estrechos posibles que contienen la probabilidad posterior requerida sobre dicho parámetro. Si la distribución posterior de probabilidad tiene forma bi-modal, el IDM está compuesto por dos intervalos disjuntos. Para los parámetros de esta investigación, calcularemos los intervalos IDM del 95% de probabilidad, salvo cuando usemos métodos de simulación, en los que no es posible su cálculo [**FIGURA 6**].

Para el cálculo de las Razones de Verosimilitud (RV) se ha hecho también análisis de referencia. Las RV son magnitudes continuas cuyo valor oscila entre 0 e ∞ , y se obtienen a partir de los valores muestrales de Sensibilidad y Especificidad:

$$RV \text{ positivos} = RV+ = S/(1-E) \text{ y } RV \text{ negativos} = RV- = (1-S)/E$$

Para nuestro análisis bayesiano, las hemos obtenido muestreando mediante técnicas de Cadena de Markov y simulación de Montecarlo con muestreo de Gibbs mediante el programa WinBUGS 14.3. Se computan los límites inferior y superior de los Intervalos de credibilidad al 95% sobre RVpositivos y RVnegativos, y se obtienen los diagramas de Caja con Patillas y los gráficos de sus densidades a posteriori.

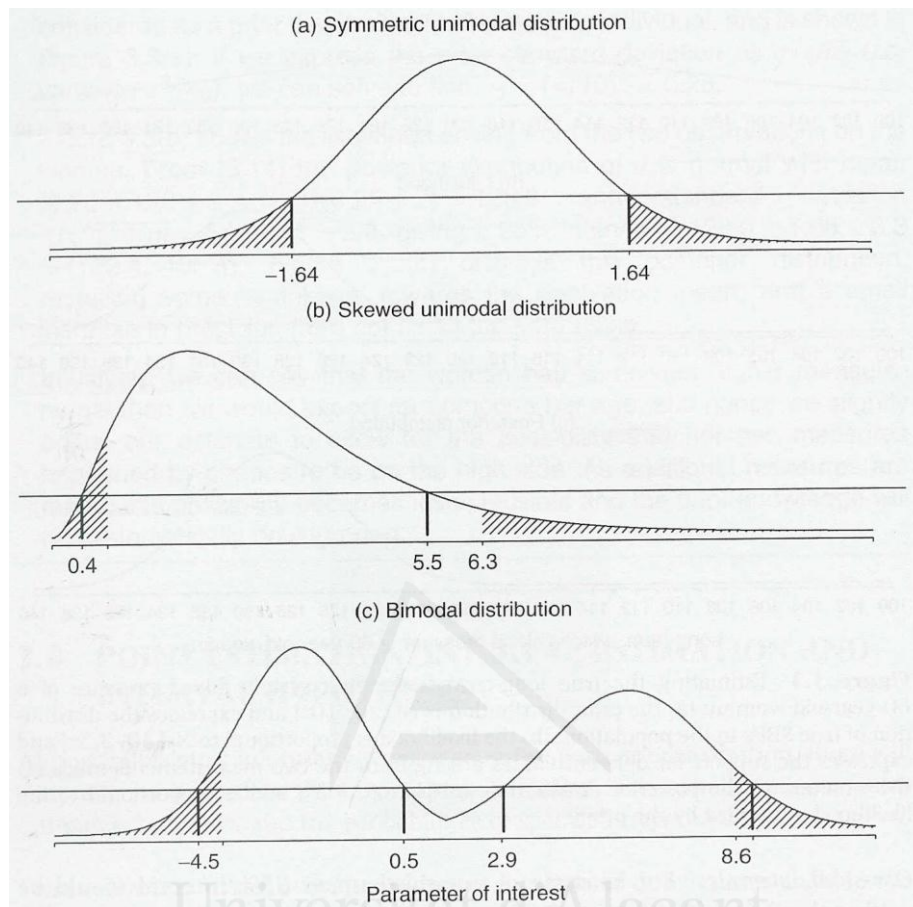


FIGURA 6: (a) muestra una distribución simétrica unimodal en la que los intervalos de credibilidad (igual área en ambas colas) y los intervalos IMD coinciden en -1.64 a 1.64. (b) representa una distribución unimodal asimétrica en la que el intervalo de credibilidad es 0.8 a 6.3, mientras que el intervalo MD es 0.4 a 5.5, considerablemente más estrecho. Entre 0.4 y 0.8 (por fuera del intervalo de credibilidad) el parámetro tiene valores mucho más probables que entre 5.5 y 6.3 (por dentro), por lo que la máxima densidad de probabilidad queda mejor representada por el IMD. (c) muestra el caso de una distribución bi-modal en la que el intervalo de igual área en ambas colas (de credibilidad) es -3.9 a 8.6, mientras que el intervalo MD consiste adecuadamente en dos segmentos disjuntos.

Puede demostrarse teóricamente⁸¹ que para cualquier tamaño muestral, los intervalos de credibilidad de referencia posterior para un parámetro dados unos datos con verosimilitud normal de tamaño muestral n , *coinciden numéricamente* con los intervalos de confianza frecuentistas calculados en base a la distribución muestral del estadístico t de Student. Esto es; la distribución de referencia posterior sobre $\ln(RV)$ puede computarse como:

$$\ln RV \approx St(\ln RV / \ln RV_{muestral}, s / \sqrt{n-1}, gl = n-1)$$

Para el cálculo de los intervalos de credibilidad y de máxima densidad y de gráfico de las distribuciones se ha utilizado el programa MatLab 7.0.1., el programa WinBUGS 14.3, accesible libremente en la dirección <http://www.mrc-bsu.cam.ac.uk/bugs>, y el programa¹⁴⁶ EpiInfo 3.1 desarrollado por el Servicio de Epidemiología de la Dirección Xeral de Saúde Pública da Consellería de Sanidade (Xunta de Galicia) en colaboración con la Unidad de Análisis de Salud y Sistemas de Información Sanitaria da Organización Panamericana de la Salud (OPS-OMS), a través da carta de entendemento existente entre a Consellería de Sanidade y la OPS-OMS, accesible libremente en Internet en la dirección <http://dxsp.sergas.es>

3.1.4.2.- Calibración del modelo:

La calibración es una medida de cuanto se acercan las probabilidades predichas por el modelo a las probabilidades reales de muerte de cada uno de los individuos. Clásicamente, la calibración del modelo se evalúa utilizando la prueba de Hosmer y Lemeshow^{125,147}. Se divide la muestra en grupos pequeños (de al menos 5 individuos) utilizando para ello los deciles de riesgo de mortalidad predicha por el modelo. Se calcula la suma de eventos predichos y la suma de eventos realmente observados dentro de cada subgrupo (vivos y muertos de cada uno de los estratos de riesgo predicho), y se determina si existen diferencias estadísticamente

significativas entre los números observados y los predichos mediante un simple test de Ji-cuadrado.

Nuestra propuesta **bayesiana** para estudiar la calibración del modelo utiliza conceptos de la teoría matemática de la información [**ANEXO IV**]. Consiste en calcular la entropía relativa (divergencia logarítmica de Kullback-Leibler¹⁴⁸), la divergencia de Jeffreys y la discrepancia intrínseca¹⁴⁹ como medidas para caracterizar **la distancia** -medida en bits de información- entre dos distribuciones de probabilidad.

De nuevo en este caso se divide a la muestra en subgrupos pequeños (estratos de riesgo) utilizando para ello los deciles de riesgo de mortalidad predicha por el modelo - de hecho la división es la misma que la realizada anteriormente con el SPSS para la prueba de Hosmer y Lemeshow - y se calcula, para toda la muestra (vivos y muertos) la Entropía Relativa entre las dos distribuciones de probabilidad de morir:

- La **estimada** por el modelo de RL
- La **realmente subyacente** al modelo probabilístico que está generando los datos de la muestra. Se han desarrollado dos versiones distintas del test de calibración, porque como estimador puntual de la probabilidad real de muerte en cada estrato de los generados por los deciles de la probabilidad predicha, se han utilizado dos tipos de estimadores bayesianos:
 - El estimador puntual bayesiano convencional: la media de la distribución Beta posterior de referencia sobre cada estrato, que, siendo r el número de vivos o muertos y n el número de niños que forma cada estrato, es igual a $(r + \frac{1}{2}) / (n + 1)$.
 - El llamado "estimador puntual intrínseco"¹⁵⁰: el que minimiza, en un análisis de decisión, la pérdida esperada intrínseca (usando la discrepancia intrínseca como función de pérdida) de actuar como si este fuera el verdadero valor. Bernardo ha mostrado que el valor exacto de este estimador puede obtenerse por

integración numérica o usando la función digamma¹⁵¹, pero una aproximación muy buena (que es la que hemos utilizado) es igual a $(r + 1/3)/(n + 2/3)$, un valor muy próximo a la mediana de la distribución Beta posterior de referencia sobre cada estrato, siendo de nuevo r el número de vivos o muertos y n el número de niños de cada estrato.

La Entropía Relativa o Divergencia Kullback-Leibler¹⁴⁸ entre las dos distribuciones de probabilidad se calcula como:

$$D(p||f) = \sum_{\text{deciles}} \sum_{\text{vivos}}^{\text{muertos}} p(x) \log_2 [p(x)/f(x)]$$

siendo $p(x)$ y $f(x)$ la probabilidad del evento, vivir o morir, en cada uno de los deciles de riesgo, para cada una de las dos distribuciones -la estimada por el modelo logístico, y la que se considera "real"- . Por argumento de continuidad, se asume para el cálculo que $0 \log 0/f = 0$ y $p \log p/0 = \infty$. Aunque suele interpretarse como una "distancia" entre distribuciones de probabilidad, por sus propiedades matemáticas no es una distancia estricta [$D(p//f) \geq 0$ con igualdad si $p(x) = f(x)$ y $D(p//f) \neq D(f//p)$]. Por ello se miden además la Divergencia de Jeffreys = $D(p//f)+D(f//p)$ y la Discrepancia intrínseca¹⁵⁰= $\text{Mín}[D(p//f), D(f//p)]$ con propiedades más adecuadas al concepto de "distancia" entre distribuciones.

El resultado global se expresa en unidades de información, y se ha escogido el bit por ser esta una medida de información generalizada y entendible por todas las personas familiarizadas con la informática doméstica. Un bit es la información proporcionada por un evento cuya probabilidad de ocurrir es $\frac{1}{2}$, y coincide con la entropía de un sistema -entendida como medida de su aleatoriedad o indeterminismo- que genera una secuencia de valores de una variable aleatoria binaria equiprobable. Así, un sistema que puede producir dos eventos equiprobales (por ejemplo el lanzamiento de una moneda perfecta) tiene una entropía de 1 bit por tirada. El test se ha realizado en un hoja de cálculo de MS-Excel construida para tal fin. El resultado es siempre un número real positivo cuyo valor es cero si y sólo si las probabilidades estimadas por el modelo y las observadas coinciden. La

interpretación del resultado, está basada en las escalas que se usan clínicamente para medir las razones de verosimilitud¹⁵², y es como sigue:

BITS	"Distancia" entre las distribuciones
0	Ambas distribuciones coinciden
0 a 0.01	Casi inexistente o despreciable
0.01 a 1	Sustancial
1 a 3.33	Grande
3.33 a 6.66	Muy grande
Más de 6.66	Enorme

3.2- RED NEURONAL ARTIFICIAL:

3.2.1.- Diseño:

Estudio observacional analítico de una cohorte retrospectiva.

3.2.2.- Muestreo y tamaño muestral:

El estudio se realiza sobre el análisis retrospectivo de la misma base de datos que la usada para ajustar el modelo logístico. Los pacientes son los mismos.

3.2.3.- Desarrollo de la Red Neuronal Artificial:

3.2.3.1.- Software empleado:

En la terminología al uso, el desarrollo de un modelo de red neuronal artificial (RN) se denomina "entrenamiento". El entrenamiento de la RN se ha llevado a cabo usando un módulo específico de desarrollo que ha sido realizado y es propiedad del Grupo de Procesado Digital de Señales (GPDS) del departamento de

Ingeniería Electrónica de la Universidad de Valencia. Su demostrada experiencia en el campo de las RN ha permitido la creación de una herramienta optimizada para la resolución automática de problemas de clasificación mediante RN. Su facilidad de uso, generalidad y la posibilidad de adecuarlo al problema, gracias al GPDS, lo dota de las condiciones idóneas para su aplicación en el presente trabajo frente a otras posibilidades comerciales de propósito similar. Programado utilizando el entorno de programación Matlab (propiedad de MathWorks), permite una gran flexibilidad a la hora de enfrentarse con cualquier tarea que requiera programación. El módulo está programado para analizar los datos del problema, dirigir el aprendizaje de las redes neuronales para una resolución óptima del problema e interpretar los resultados para facilitar la comprensión del modelo final.

Para el manejo del módulo, únicamente es necesario proporcionar el nombre del fichero que contiene los datos del problema. Deben respetarse ciertos requisitos habituales para no entrar en las excesivas opcionalidades del programa. El fichero de texto debe estar organizado en columnas, contener en su primera línea los nombres de las variables independientes y en último lugar la variable dependiente (codificada como uno y cero).

En nuestro caso, las redes se han entrenado sobre un ordenador compatible PC con microprocesador AMD Athlon XP 1500, 1,5 GB de RAM y ejecutando el sistema operativo Linux 2.4.18. Sobre este sistema operativo corría un servidor de Matlab 5.1 y un compilador Matcom, que mejoraba la rapidez en la ejecución de los ficheros .m.

El resultado del proceso, llevado a cabo por el módulo, consiste en información detallada del modelo neuronal que mejor resuelve el problema, así como datos sobre su rendimiento (matriz de confusión, Se, Sp, VPP, VPN, RVP, RVN), análisis de su capacidad discriminatoria (curvas ROC) y estudio de la importancia de las variables del problema (análisis de sensibilidad).

3.2.3.2.- Selección del mejor modelo de RN:

Para entrenar una RN son necesarias dos fases sucesivas: entrenamiento propiamente dicho y evaluación del mejor modelo. El entrenamiento de RN consiste básicamente en un proceso exhaustivo de búsqueda de la red óptima. Partiendo de una arquitectura basada en el Perceptrón Multicapa de una capa oculta completamente conectada, con función de activación tangente hiperbólica²⁵⁴ y con un número de neuronas en la capa de entrada y salida fijadas por el problema, el problema se basa en encontrar el número óptimo de neuronas en la capa oculta así como el valor de las conexiones de la red.

Con este fin, el módulo programa una batería de entrenamientos para cada una de las posibilidades. Empezando por una neurona oculta, se repite el proceso hasta alcanzar el máximo permitido. La experiencia dicta que no se deben permitir más grados de libertad en un modelo matemático que muestras disponibles si se quiere asegurar un grado de generalidad en los resultados, es decir, disponer de estructuras neuronales con más conexiones que datos disponibles.

Cada fase de entrenamiento se ha realizado mediante el algoritmo de *retropropagación del error*²⁵⁴. Básicamente consiste en encontrar los valores de las conexiones de la red neuronal, inicialmente aleatorias, tales que obtengan el mínimo error cuadrático medio calculado como la diferencia entre la salida de la red y la respuesta deseada. La implementación del algoritmo en el módulo únicamente requiere de un parámetro, denominado constante de adaptación, que regula el cambio permitido en cada iteración. Este se selecciona automáticamente a un valor de 0.5 y se decrementa de manera lineal con las iteraciones.

La finalización del entrenamiento se ha realizado mediante la técnica de validación cruzada. Los datos del problema son divididos aleatoriamente en dos grupos uno de ellos con el 25% para evaluar constantemente el grado de sobreajuste o memorización a los datos utilizados para el aprendizaje. En el momento que se detecte un crecimiento en el error calculado para los datos de contraste se detendrá el entrenamiento.

Finalmente el propio módulo evalúa los diferentes modelos obtenidos y selecciona el mejor de acuerdo a un criterio establecido por el usuario. En este caso, se ha elegido una valoración dependiente de los indicadores de la matriz de confusión. Se descartarán todos los modelos con Sensibilidad y Especificidad inferiores a 0,7 por considerarlos insuficientes para utilizarlos como una prueba diagnóstica médica. Entre los modelos que superen el criterio se valorará el valor más alto en la suma de Sensibilidad y Especificidad, tanto para los datos de entrenamiento como para los de validación. Por último, se plasmarán en forma de histogramas la categoría de ocurrencias predichas en casa intervalo de salidas obtenidas con las RN.

3.2.4.- Evaluación de la exactitud diagnóstica del modelo:

La calidad de la RN como método de predicción se ha evaluado midiendo también su poder de discriminación y su calibración, exactamente del mismo modo como se ha hecho con el modelo de RL.

3.2.4.1.- Capacidad discriminante:

Las medidas usadas para evaluar la capacidad discriminante son, de nuevo, las medidas clásicas: la Sensibilidad, la Especificidad, las Razones de Verosimilitud para positivos y para negativos, la Odds Ratio Diagnóstica y el área bajo la curva ROC (AUC). Su cálculo se ha realizado con el mismo procedimiento usado para la RL (ver arriba).

La evaluación **bayesiana** de la capacidad discriminante del modelo de RN ha sido idéntica a la de la RL: se ha utilizado el teorema de Bayes y se ha realizado un análisis de referencia para estimar los Intervalos de Credibilidad al 95% de la Sensibilidad, la Especificidad y el AUC, con sus distribuciones a posteriori de referencia. De nuevo se ha partido de la distribución a priori menos informativa para el problema: la Beta (0'5, 0'5)^{57,142}. Se han utilizado los programas MatLab 7.0.1. , WinBUGS 14.3 y EpiInfo 3.1.

3.2.4.2.- Calibración del modelo:

La calibración clásica se ha evaluado con la prueba de Hosmer y Lemeshow^{125,147}, con el mismo procedimiento que el utilizado para la RL.

Nuestra propuesta **bayesiana** para estudiar la calibración del modelo es idéntica también a la detallada con la RL, desarrollada por nosotros desde conceptos de la teoría matemática de la información [**ANEXO IV**]. Consiste en calcular medidas para caracterizar *la distancia* -medida en bits de información- entre dos distribuciones de probabilidad.

De nuevo en este caso se divide a la muestra en subgrupos pequeños (estratos de riesgo) utilizando para ello los deciles de riesgo de mortalidad predicha por el modelo - de hecho la división es la misma que la realizada anteriormente con el SPSS para la prueba de Hosmer y Lemeshow - y se calcula, para toda la muestra (vivos y muertos) la Entropía Relativa, la Divergencia de Jeffreys y la Discrepancia intrínseca de Bernardo, entre las dos distribuciones de probabilidad de morir:

- La **estimada** por el modelo de RN
- La **realmente subyacente** al modelo probabilístico que está generando los datos de la muestra. Se han desarrollado dos versiones distintas que han utilizado dos tipos de estimadores bayesianos:
 - El estimador puntual bayesiano convencional: $(r + \frac{1}{2}) / (n + 1)$.
 - El llamado "estimador puntual intrínseco": $(r + 1/3) / (n + 2/3)$.

El test se ha realizado en la misma hoja de cálculo de MS-Excel usada para la RL, construida por nosotros para tal fin. El resultado global se expresa en bits. La interpretación del resultado, está basada en la misma escala que la usada con la RL (pág 89).

4.- 2ª FASE: VALIDACIÓN DE AMBOS TESTS

Para obtener un estimador insesgado del poder de discriminación y de la calibración de un modelo -esto es, para conocer el comportamiento del modelo como método de predicción en futuras muestras (inferencia a nivel poblacional)- los índices de rendimiento diagnóstico deben calcularse sobre una muestra de datos no utilizados en el desarrollo del modelo (validación externa)^{153,154}. Esto será lo que realicemos en esta fase, la llamada "validación propiamente pre-dictiva" de los modelos.

4.1.- DISEÑO:

Estudio observacional analítico de una cohorte retrospectiva distinta, pero clínicamente superponible.

4.2.- MUESTREO:

El estudio de validación se realiza sobre el análisis retrospectivo de la base de datos de los pacientes ingresados en la UCI Pediátrica del Hospital Infantil "La Fe" durante los años 1994 y 1995. Los pacientes son los mismos para validar ambos modelos, la RL y la RN.

4.3.- VALIDACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA:

Para cada paciente de la cohorte de validación se ha calculado la probabilidad de muerte predicha por el modelo de RL ajustado en la primera fase de desarrollo. Esto es, se ha calculado:

$$P(\text{morir} / \text{RL}) = \frac{1}{1 + e^{-(-6.016 + 0.146 * \text{PRISM} 24)}}$$

Para clasificar los pacientes en muertos o vivos para el modelo, se ha usado el punto de corte probabilidad pronosticada ≥ 0.0111 .

4.4.- VALIDACIÓN DEL MODELO DE RED NEURONAL ARTIFICIAL:

La RN se ha validado también sobre un ordenador compatible PC con microprocesador AMD Athlon XP 1500, 1,5 GB de RAM y ejecutando el sistema operativo Linux 2.4.18. Sobre este sistema operativo corre un servidor de Matlab 5.1 y un compilador Matcom, que mejoraba la rapidez en la ejecución de los ficheros .m.

Para la fase de validación se ha utilizado la misma RN desarrollada en la fase anterior, pero ahora se le han enseñados los datos de la nueva cohorte de validación. Para la clasificación final de los pacientes, se ha usado el mismo punto de corte: probabilidad pronosticada ≥ 0.0076806 .

4.5.- EVALUACIÓN DE LA EXACTITUD DIAGNÓSTICA DE AMBOS MODELOS:

La calidad de ambos modelos (RL y RN) como métodos de predicción se ha evaluado midiendo también sus poderes de discriminación y su calibración, pero

sobre la nueva muestra. Tanto en la versión *clásica* como en nuestra propuesta *bayesiana*, se ha seguido la misma metodología explicada para la fase de desarrollo del modelo (ver arriba).

4.5.1.- Evaluación de la capacidad discriminante:

La capacidad de discriminación se ha evaluado con el estudio de la curva ROC y de los índices de rendimiento diagnóstico, tanto en su versión clásica como en nuestra propuesta bayesiana.

4.5.2.- Evaluación de la Calibración:

La calibración clásica se ha estudiado de manera idéntica a lo que se ha hecho en la fase de desarrollo. En esta caso, por simplicidad, tanto la realización del test de Hosmer-Lemeshow como el estudio de las entropías informativas se ha hecho, no mediante la hoja de cálculo de MS Excel, sino desarrollando un archivo de programación computacional mediante el programa MatLab7. Se exponen sólo la distribución en deciles de riesgo y los resultados numéricos de los tests.

5.- 3ª FASE: COMPARACIÓN DE LA EXACTITUD DIAGNÓSTICA DE AMBOS TEST

5.1.- COMPARACIÓN DE LA CAPACIDAD DISCRIMINANTE

Es natural que la comparación de la eficacia de dos o más pruebas diagnósticas para detectar una enfermedad o proceso patológico dado, pueda hacerse sobre la base de comparar sus índices de exactitud diagnóstico: los valores de Sensibilidad, Especificidad y las Razones de verosimilitud. También

podría utilizarse el Odds Ratio Diagnóstico (ORD). Pero cuando se trata de pruebas diagnósticas con resultado cuantitativo, la comparación de las curvas ROC correspondientes resulta el modo más natural de determinar cuál de las pruebas es más eficaz, ya que el valor de los índices de rendimiento diagnóstico de pruebas de este tipo depende del punto de corte que se elija.

Teniendo en cuenta lo que se ha visto hasta ahora se comprende que la curva ROC que tenga el área mayor será la que corresponde a la prueba más exacta. Para comparar ambos modelos se usan las curvas ROC obtenidas sobre los datos de la cohorte de validación. Las curvas ROC se han trazado con el programa SPSS.

5.1.1.- Evaluación Clásica:

Las comparaciones analíticas de las áreas bajo la curva ROC (AUC) de ambos métodos se han hecho con Epidat 3.1. Este programa puede hacer comparaciones de AUCs *cuando las curvas son correlacionadas*, esto es, construidas con los mismos pacientes. El programa utiliza el método no paramétrico de DeLong¹⁴¹ para calcular la significación estadística de la diferencia en las áreas, que tiene como hipótesis nula la de igualdad de las AUCs.

No puede calcularse el IC95% de la diferencia en las AUCs.

5.1.2.- Propuesta de Evaluación Bayesiana:

Nuestra propuesta de comparación bayesiana de las AUCs incluye varias opciones, pero se basan todas en calcular si el Intervalo de Credibilidad al 95% de la distribución posterior sobre la diferencia entre proporciones incluye al valor cero, que representa la igualdad entre las proporciones.

a) Utilizando la aproximación Normal a la distribución Beta:

Cuando se utilizan densidades posteriores, el cálculo del intervalo de credibilidad debe realizarse calculando áreas bajo la distribución posterior. Ya que

se ha modelizado el AUC como una proporción, y en la fase de validación se ha obtenido la distribución posterior sobre el AUC de según cada uno de los dos métodos, procedemos ahora a encontrar e integrar la distribución posterior sobre la diferencia en AUCs: la distribución posterior sobre una diferencia de proporciones.

El principal problema radica en que para calcular directamente áreas bajo la distribución posterior de la diferencia entre dos distribuciones beta se precisa de métodos complicados de cálculo. Para simplificar el trabajo, nuestra propuesta se basa en el método descrito por Berry^{155,156} y consiste en aplicar una propiedad de la familia de densidades beta: cuando el tamaño muestral es grande y los parámetros a y b de una distribución beta son razonablemente grandes, la distribución beta se vuelve una distribución simétrica que puede aproximarse con una distribución normal con las mismas media y varianza. Y aunque la diferencia de dos distribuciones beta no tiene que seguir necesariamente una distribución beta, se sabe que la diferencia entre dos distribuciones normales si que sigue una distribución normal con media la diferencia de las medias y varianza la suma de las varianzas. Así pues, usando la aproximación normal a la distribución beta, podemos calcular el Intervalo de Credibilidad bajo la posterior de la diferencia. Este cálculo se ha realizado con el programa MatLab 7.0.

b) Utilizando la simulación con métodos de Monte Carlo:

Otra aproximación es utilizar técnicas de simulación con métodos de cadena de Markov y simulación de Monte Carlo. En nuestro caso se han realizado con el programa EpiDat 3.1: se han obtenido las distribuciones beta posteriores sobre las AUCs según ambos métodos y se han realizado 10000 simulaciones de muestreo sobre cada una de ellas, y representado la distribución empírica posterior de la diferencia entre los valores obtenidos. Se indican los percentiles relevantes.

c) Utilizando el método gráfico discreto de la parrilla de Berry^{155,156}:

Usando las distribuciones de referencia no informativas Beta (1/2, 1/2), y utilizando el AUC como una probabilidad cuyos datos muestrales se modelizan con la distribución Binomial, podemos calcular las distribuciones posteriores de referencia para las dos AUC de cada test diagnóstico, la de la RL y la de la RN. El método de Berry representa los resultados en una parrilla discreta (en este caso de 100 casillas para cada lado), y asume que ambos métodos diagnósticos son independientes: la probabilidad del suceso conjunto se calcula mediante *el producto* de sus probabilidades. Los valores de la posterior sobre AUC de RL corresponden a las filas de la parrilla, y los de la RN corresponden a las columnas. Cada cuadro de la parrilla representa un par (AUC/RL, AUC/RN) y en la tercera dimensión se representa la probabilidad conjunta (el producto) de ambas probabilidades posteriores.

Dispuestos así los datos, la diagonal de la parrilla tiene un significado especial: en ella se sitúan las casillas que representan a los pares para los que $AUC/RL = AUC/RN$, y la altura en la tercera dimensión de esta diagonal representa la situación en la que ambas AUC son idénticas. Si la curva acampanada queda por debajo de la diagonal, la AUC de la RL es superior. Si la curva queda por encima, la RN es superior.

La probabilidad de que la diferencia en ambas AUC sea igual a cero ($difAUC = AUC/RL - AUC/RN = 0$), puede calcularse sumando los valores asignados a las casillas que ocupan la diagonal (esto es aquellas en que $AUC/RL = AUC/RN$). De igual modo, la probabilidad de que la diferencia entre ambas AUC sea de 0.25 ($difAUC = 0.25$) se calcula sumando las probabilidades asignadas a las casillas en las que $AUC/RL - AUC/RN = 0.25$. Así podemos calcular las probabilidades posteriores sobre todos los valores discretos de la variable $difAUC$ entre -1 y 1 que queden representados por las $100 \times 100 = 10000$ casillas de nuestra parrilla.

Los cálculos se han hecho con el programa MatLab 7.0, pero la representación gráfica se ha hecho con una parrilla de 10 casillas por lado en una hoja de cálculo de Excel construida a propósito para el análisis.

5.2.- COMPARACIÓN MEDIANTE INCORPORACIÓN AL PROCESO DIAGNÓSTICO: ANÁLISIS DE DECISIÓN:

Desde el punto de vista clínico, la manera más útil de comparar ambos métodos diagnósticos es incorporarlos a un análisis formal de decisiones. Para ello, se ha realizado un análisis de decisión para intentar contestar a la pregunta de cuál de los dos métodos es preferible para usarlo clínicamente en la información a los padres de los niños que ingresan en la UCIP. Ya que el objetivo es eminentemente práctico, como instrumento de mejora en la exactitud de la estimación del pronóstico vital de los niños que sirva de ayuda para informar a sus padres después de las primeras 24 horas del ingreso en intensivos, el análisis se hace desde la perspectiva del médico que va a utilizar el test.

Para representar adecuadamente esta perspectiva, hemos realizado una encuesta entre los médicos de la UCI pediátrica del Hospital Inafantil de La Fe, que son los usuarios potenciales de los test diagnósticos. Todos los médicos coincidieron en asignar la máxima utilidad a los Verdaderos negativos y la mínima a los Falsos negativos. Utilizando el clásico método del "juego estándar", cada médico determinó posteriormente las utilidades particulares que asignaba a cada una de las dos trayectorias restantes. Finalmente, se asumieron por consenso las siguientes utilidades finales:

- Verdaderos negativos: Máxima utilidad. Utilidad = 1
 - Corresponde al acierto pleno, porque se identifican exactamente los niños que no morirán. Se da correctamente a los padres una buena noticia.
- Falsos negativos: Mínima utilidad. Utilidad = 0

- Desde el punto de vista clínico, corresponde a la peor posibilidad. Se les dice a los padres que sus hijos no morirán y sí lo harán. No sólo se produce un diagnóstico erróneo, sino que además se crean en los padres falsas expectativas de mejoría que pueden traer consecuencias funestas: se pierde credibilidad deteriorando la relación médico-familia, y se puede interferir en la elaboración del duelo.
- Verdaderos positivos: Utilidad = 0,8
 - Corresponde también a un acierto, pues se identifican exactamente los niños que morirán. Pero se consideró de menor utilidad que la de los Verdaderos negativos, pues se dá a los padres (con exactitud) una muy mala noticia.
- Falsos positivos: Utilidad = 0,5
 - Constituye un fallo en el diagnóstico, pero el resultado final de esta trayectoria no tiene la peor utilidad, porque aunque el diagnóstico es erróneo, el niño acaba finalmente viviendo. Los padres suelen perdonar este tipo de errores.

Dada la naturaleza del análisis, no se ha tenido en cuenta ningún horizonte temporal ni se han aplicado tasas de descuento en las utilidades. Para tomar en consideración la incertidumbre en la estimación de la exactitud de los test y en la toma de las decisiones clínicas que se derivan del uso de los test, se ha hecho un análisis de sensibilidad de la decisión dominante en función de la prevalencia previa: de la estimación de mortalidad previa a la aplicación del test.



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
VI. - Resultados.
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

1.- 1ª FASE: DESARROLLO DE LOS TESTS

1.1.- MODELO PREDICTIVO DE REGRESIÓN LOGÍSTICA:

La muestra estuvo formada por los datos de un total de 2032 niños, entre los que se observaron un total de 43 muertos. El modelo final de regresión logística se ajustó sólo con los 2008 niños de los que se dispuso de todos los datos, y para la predicción de mortalidad en la muestra, además de la constante, sólo estuvo formado por la puntuación PRISM, tal y como se detalla a continuación:

Pruebas omnibus sobre los coeficientes del modelo

	Chi-cuadrado	gl	Sig.
Paso	181,343	1	,000
Bloque	181,343	1	,000
Modelo	181,343	1	,000

Resumen del modelo

	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
	234,288	,086	,462

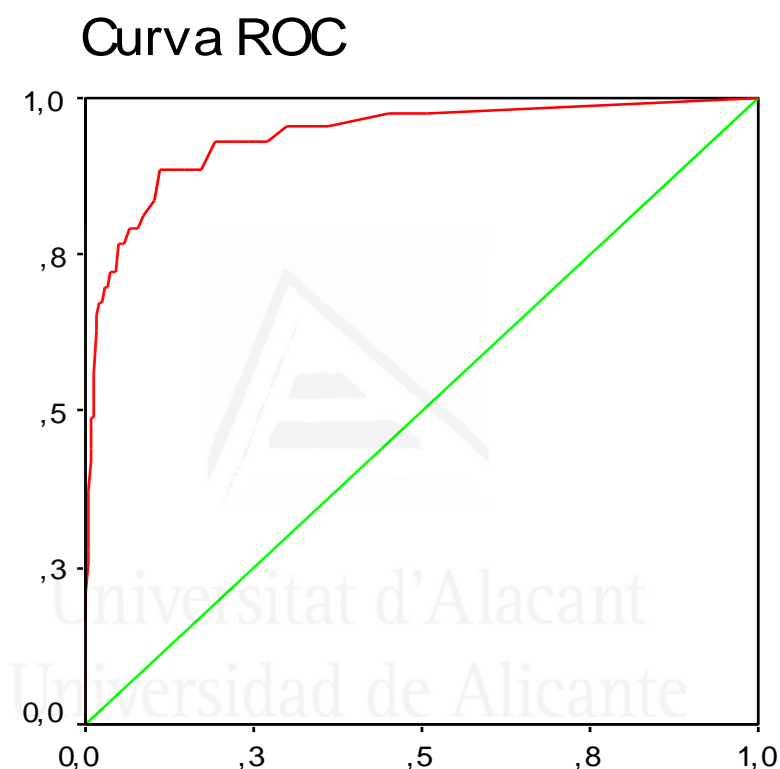
Variables en la ecuación

	B	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
				Inf erior	Superior
PRIMS24	,146	,000	1,157	1,129	1,186
Constante	-6,016	,000	,002		

1.1.1.- Evaluación clásica de la exactitud diagnóstica:

1.1.1.1.- Capacidad discriminante:

a) Estudio de la Curva ROC:



1 - Especificidad (%Falsos Positivos)

Los segmentos diagonales son producidos por los empates.

El punto de corte óptimo para el máximo rendimiento diagnóstico es una probabilidad pronosticada $\geq 0'0111$, que dá al modelo una Sensibilidad de 0'884 y una Especificidad de 0'829.

Área bajo la curva

Variables resultado de contraste: Probabilidad pronosticada

Área	Error típ. ^a	Sig. asintótica ^b	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
,939	,021	,000	,898	,980

La variable (o variables) de resultado de contraste: Probabilidad pronosticada tiene al menos un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Los estadísticos pueden estar sesgados .

- a. Bajo el supuesto no paramétrico
- b. Hipótesis nula: área verdadera = 0,5

b) Índices de exactitud diagnóstica:

Pronosticado RL	Observado		
	Muerto	Vivo	
Positivo (Muerto)	38	336	374
Negativo (Vivo)	5	1629	1634
	43	1965	2008

		IC 95%
Sensibilidad	88,4%	75,5% a 94,9%
Especificidad	82,9%	81,2% a 84,5%
Valor predictivo positivo	10,2%	7,5% a 13,6%
Valor predictivo negativo	99,7%	99,3% a 99,9%
Proporción de falsos positivos	17,1%	15,5% a 18,8%
Proporción de falsos negativos	11,6%	5,1% a 24,5%
Acierto	83,0%	81,3% a 84,6%
Odds ratio diagnóstica	36,85	14,40 a 94,31
Índice J de Youden	0,7	
Razón de Verosimilitudes + [LR(+)]	5,17	4,47 a 5,98
Razón de Verosimilitudes - [LR(-)]	0,14	0,06 a 0,32
Probabilidad pre-prueba (Prevalencia)	2,1%	

1.1.1.2.- Calibración del modelo:

Tabla de contingencias para la prueba de Hosmer y Lemeshow

		Exitus (en ese ingreso) = Vivo		Exitus (en ese ingreso) = Muerto		Total
		Observado	Esperado	Observado	Esperado	
Estratos	1	963	961,654	1	2,346	964
(deciles	2	120	119,620	0	,380	120
de riesgo)	3	174	174,238	1	,762	175
	4	177	177,056	1	,944	178
	5	195	195,331	2	1,669	197
	6	204	204,289	4	3,711	208
	7	132	132,813	34	33,187	166

Prueba de Hosmer y Lemeshow

	Ji-cuadrado	gl	Sig.
	1,348	5	,930

No se puede descartar la hipótesis nula de "bondad de ajuste"

1.1.2.- Evaluación bayesiana de la exactitud diagnóstica:

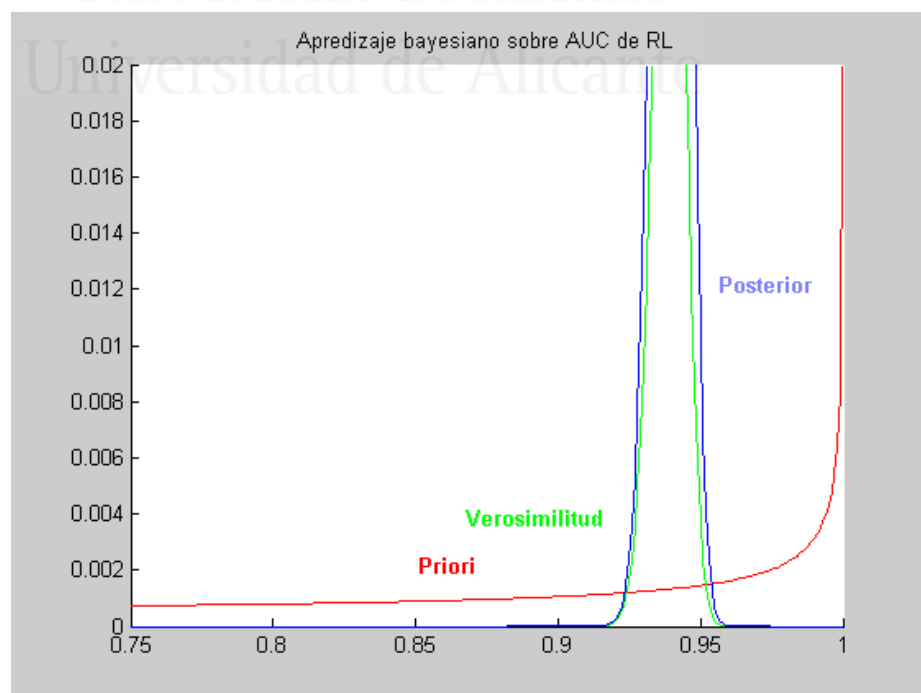
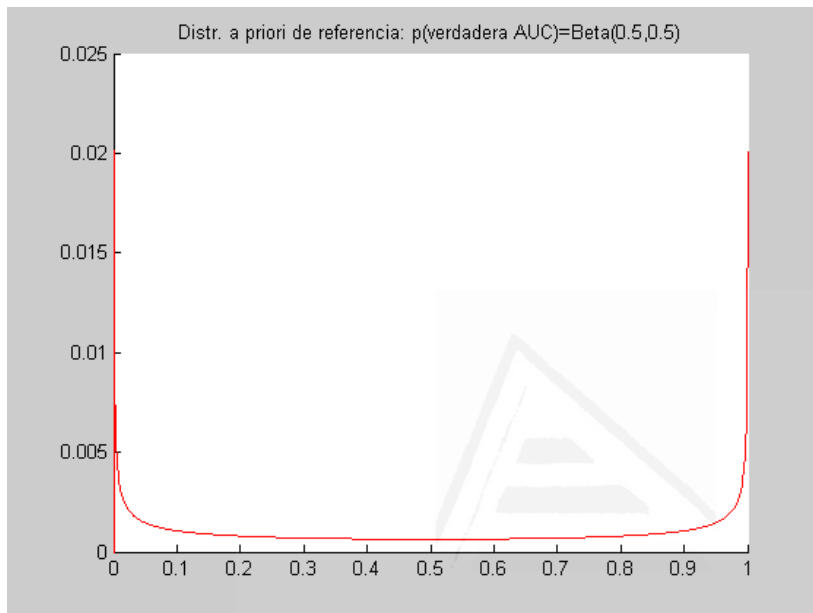
1.1.2.1.- Capacidad discriminante:

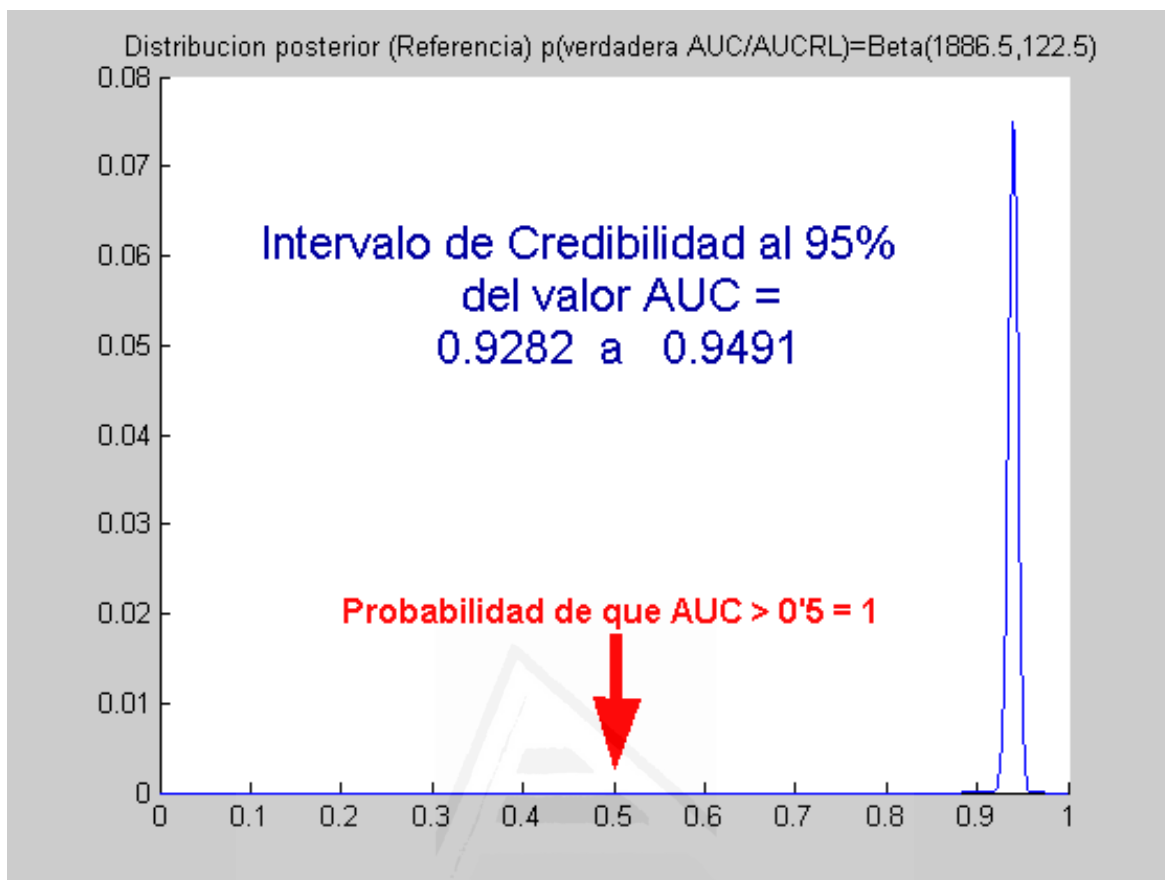
a) Estudio bayesiano de la Curva ROC:

Se considera que la AUC es una probabilidad, por lo que se modeliza con la distribución BETA. Para el análisis bayesiano de referencia, se toma como distribución a priori sobre el parámetro "verdadero valor de la AUC" la **Beta (1/2, 1/2)**. Por el estudio clásico y mediante el método trapezoidal, se sabe que el área bajo curva ROC es 0.939, y que el tamaño muestral es $n = 2008$, por lo que la función de verosimilitud conjugada de la Beta que modeliza el experimento es la Binomial ($r=1886, n=2008, AUC$).

Aplicando el teorema de Bayes, se obtiene que la distribución posterior sobre el parámetro "verdadero valor de la AUC" es la distribución **Beta (1886.5, 122.5)**, cuya media es **AUC = 0.939**. Integrándola, podemos calcular que tenemos

una probabilidad del 95% de que el "verdadero valor de la AUC" se sitúe entre 0.9282 y 0.9491 (Intervalo de Credibilidad al 95%). Y además, que la probabilidad de que el "verdadero valor de la AUC" sea mayor que 0.5 -esto es, que el modelo de regresión logística clasifique mejor que el azar- es, después de aprender de los datos del experimento, prácticamente 1.





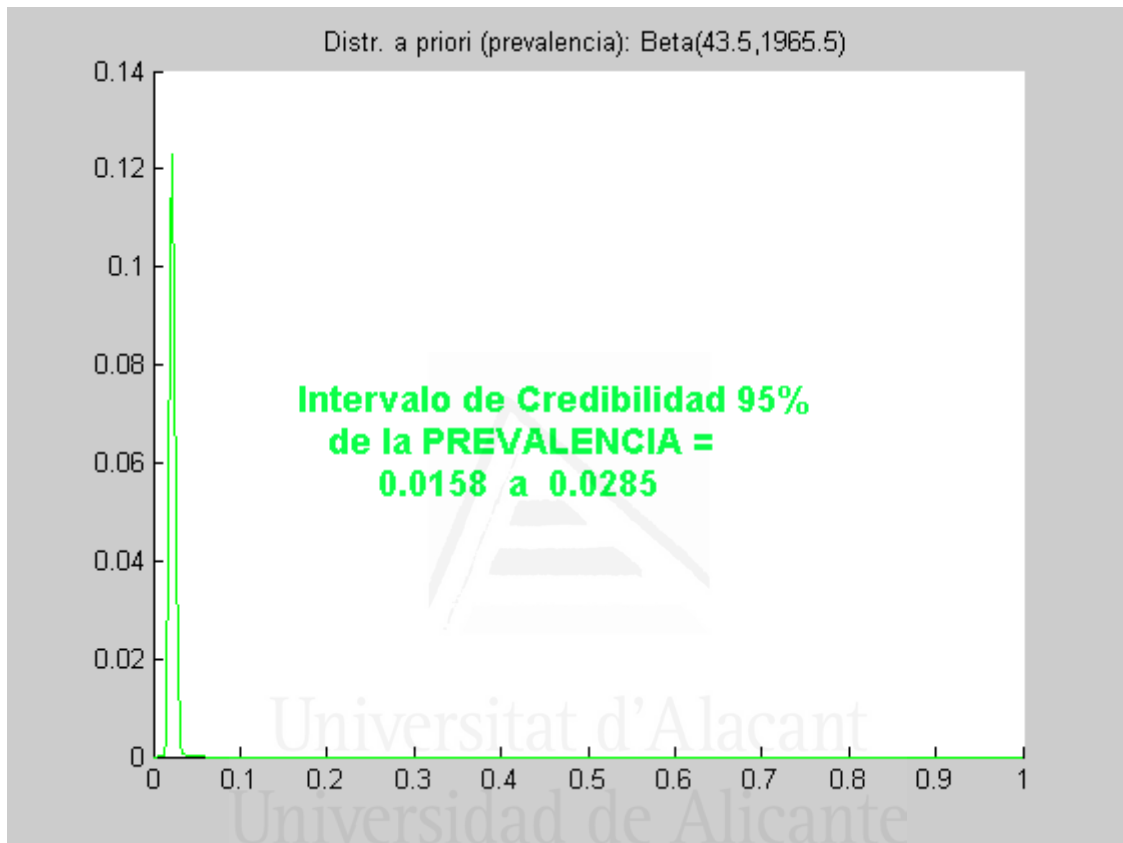
b) Índices bayesianos de exactitud diagnóstica:

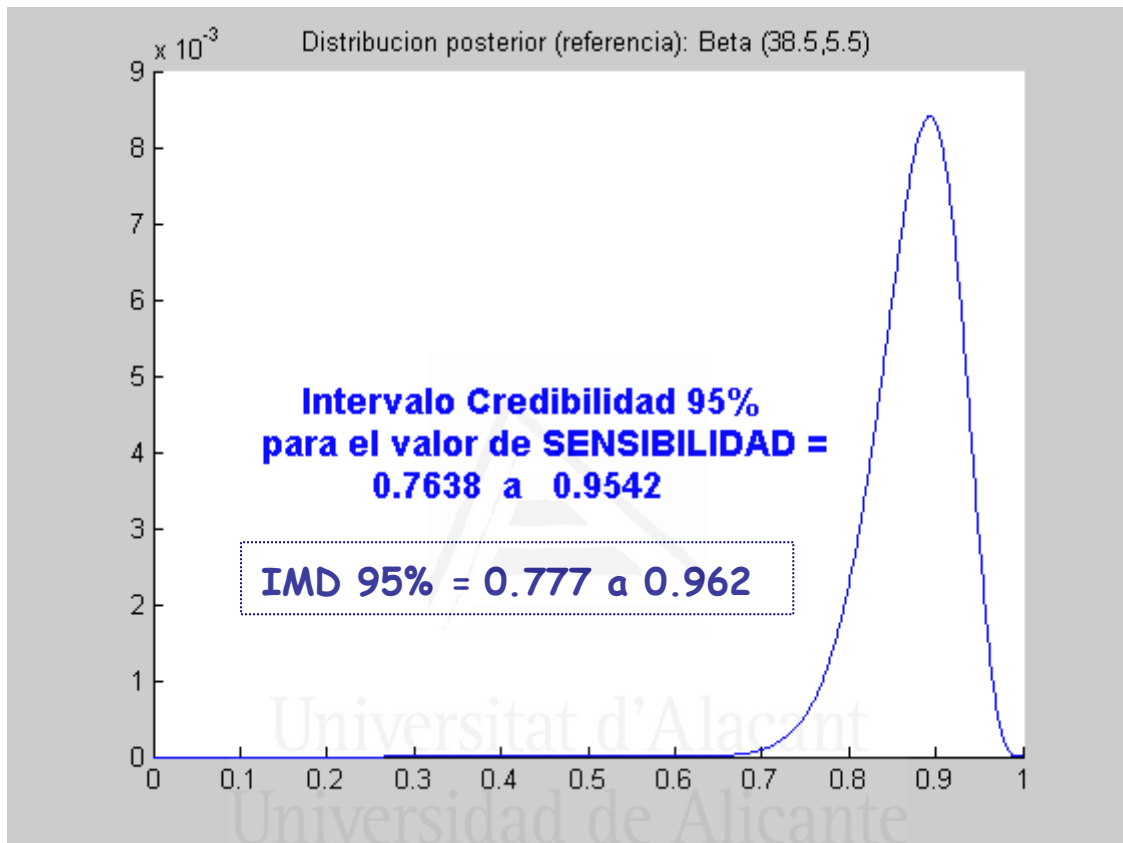
Tanto la Prevalencia como la Sensibilidad, la Especificidad y los Valores Predictivos son probabilidades, por lo que también se modelizan con la distribución BETA. Las Razones de Verosimilitud se modelizan con la distribución NORMAL, usando la transformación logarítmica. Para el análisis bayesiano de referencia, se toma como distribución a priori la **Beta (1/2, 1/2)**. En la tabla siguiente se detallan sus distribuciones de verosimilitud, sus distribuciones posteriores, y los Intervalos de Credibilidad y de Máxima Densidad al 95%. Como en el caso anterior, los resultados se han obtenido aplicando el Teorema de Bayes.

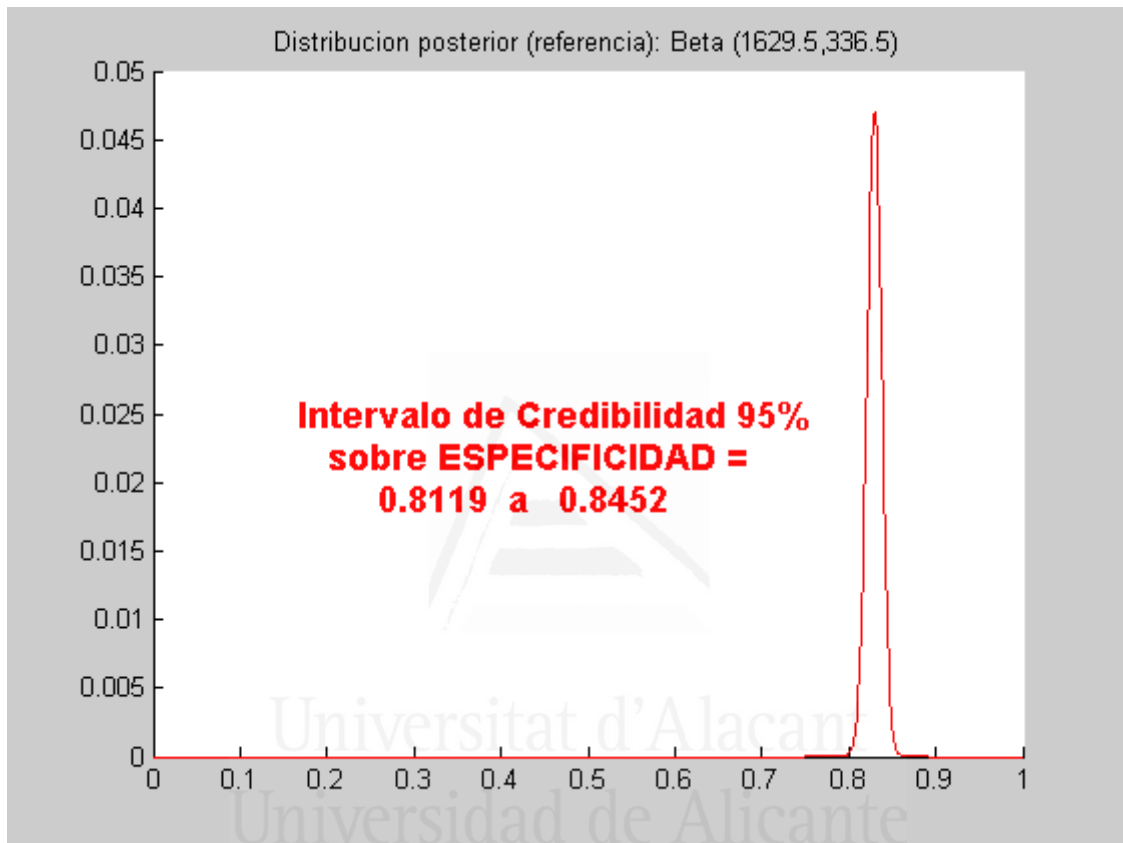
	Observado		
	Muerto	Vivo	
Pronosticado RL			
Positivo (Muerto)	38	336	374
Negativo (Vivo)	5	1629	1634
	43	1965	2008

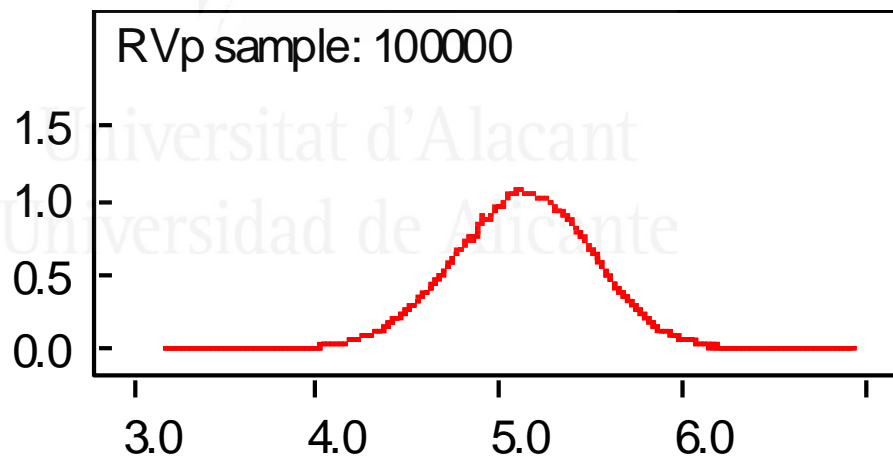
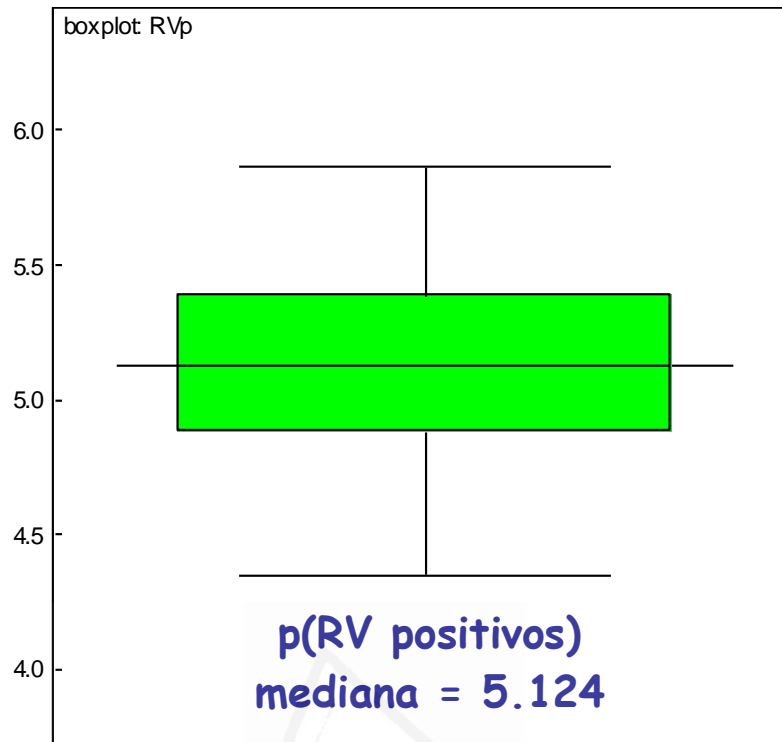
	Priori	Verosimilitud	Posteriori
Prevalencia	Be(1/2,1/2)	Bi(r=43,n=2008,P)	Be(43.5, 1965.5)
Sensibilidad		Bi(r=38,n=43,S)	Be(38.5, 5.5)
Especificidad		Bi(r=1629,n=1965,E)	Be(1629.5, 336.5)
VP Positivos		Bi(r=38,n=374,VPP)	Be(38.5, 336.5)
VP Negativos		Bi(r=1629,n=1634,VPN)	Be(1629.5, 5.5)
Raz Veros +	RVpos = Sens/(1 - Esp)		
Raz Veros -	RVneg = (1 - Sens)/Esp		

	Media (Post)	Int Credib 95%	IMD 95%
Prevalencia	2.165 %	1.58 a 2.85 %	1.58 a 2.85 %
Sensibilidad	87.5 %	76.38 a 95.42 %	77.7 a 96.2 %
Especificidad	82.884 %	81.19 a 84.52 %	81.19 a 84.52 %
VP Positivos	10.266 %	7.4 a 13.5 %	7.3 a 13.4 %
VP Negativos	99.664 %	99.3 a 99.9 %	99.3 a 99.9 %
Raz Veros +	5.124	4.347 a 5.863	
Raz Veros -	0.151	0.054 a 0.291	





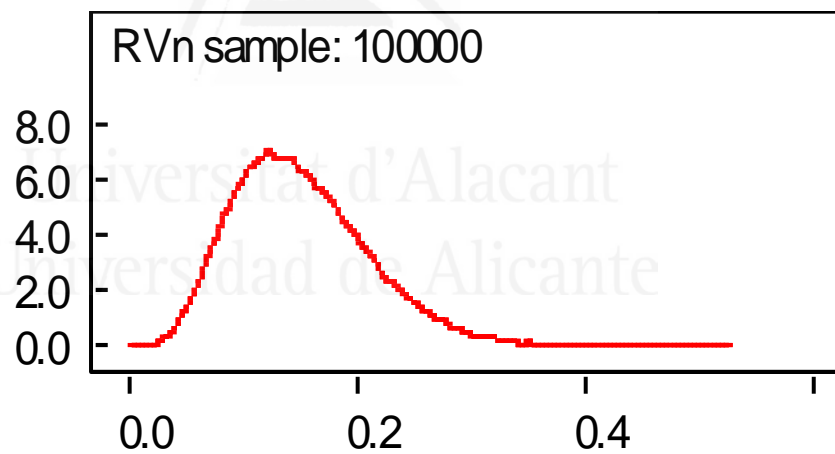
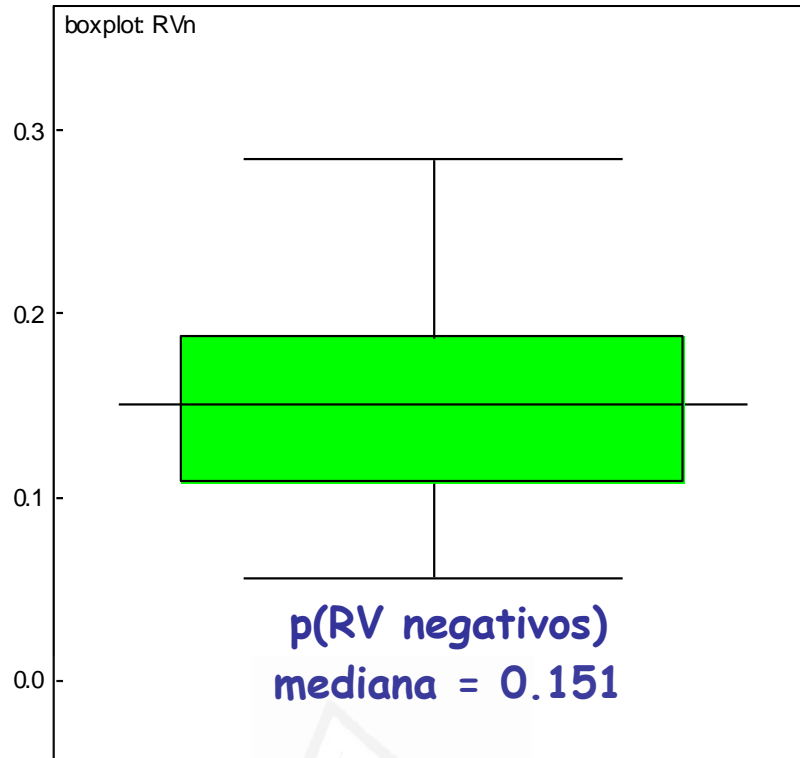




ICredib 95% RV+ = 4.347 a 5.863

Si consideramos que una RV+ es fuerte cuando es superior a 5 (TABLA 3 y TABLA 4, pag.43), puede calcularse la probabilidad de que la Regresión Logística presente una $RV+ > 5$ y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV+ > 5) = 0.6354$$



ICredib 95% RV- = 0.054 a 0.291

Si consideramos que una RV- es fuerte cuando es inferior a 1/5 (TABLA 3 y TABLA 4, pag.43), puede calcularse la probabilidad de que la Regresión Logística presente una $RV- < 1/5$. y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV- < 1/5) = 0.8043$$

1.1.2.2.- Calibración del modelo:



Universitat d'Alacant
Universidad de Alicante

Muerte en UCIP estimada con el índice "PRISM":

Comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística, y una red neuronal artificial. Una propuesta bayesiana

A) Usando como estimador puntual por estrato el ESTIMADOR BAYESIANO CONVENCIONAL

		Exitus (en ese ingreso) = Vivo				Exitus (en ese ingreso) = Muerto			
"Estratos" de Riesgo Predicho	Nº niños	Nº Observado Vivos	p(vivir)	Nº Esperado Vivos	f(vivir)	Nº Observado Muertos	p(morir)	Nº Esperado Muertos	f(morir)
1	964	963	0,998445596	961,654	0,997050777	1	0,001554404	2,346	0,002949223
2	120	120	0,995867769	119,62	0,992727273	0	0,004149378	0,38	0,007272727
3	175	174	0,991477273	174,238	0,992829545	1	0,008547009	0,762	0,007170455
4	178	177	0,991620112	177,056	0,991932961	1	0,008403361	0,944	0,008067039
5	197	195	0,987373737	195,331	0,989045455	2	0,012658228	1,669	0,010954545
6	208	204	0,9784689	204,289	0,979851675	4	0,021582734	3,711	0,020148325
7	166	132	0,793413174	132,813	0,798281437	34	0,207207207	33,187	0,201718563
8									
9									
10									
Total	2008	1965		1965,001		43		42,999	

"Estratos"	p(v)	p(v)*log2[1/p(v)]	f(v)	f(v)*log2[1/f(v)]	p(v)*log2[p(v)/f(v)]	f(v)*log2[f(v)/p(v)]	p(m)	p(m)*log2[1/p(m)]	f(morir)	f(m)*log2[1/f(m)]	p(m)*log2[p(m)/f(m)]	f(m)*log2[f(m)/p(m)]
1	0,998445596	0,002240787	0,997050777	0,004248549	0,002013705	-0,002010892	0,001554404	0,014501693	0,002949223	0,024789543	-0,001436228	0,002725003
2	0,995867769	0,005949216	0,992727273	0,010454081	0,004537937	-0,004523626	0,004149378	0,032833566	0,007272727	0,051660275	-0,003359342	0,005888011
3	0,991477273	0,01224315	0,992829545	0,010307602	-0,001949588	0,001952247	0,008547009	0,058721066	0,007170455	0,051080308	0,002165427	-0,00181667
4	0,991620112	0,012038826	0,991932961	0,011591207	-0,000451275	0,000451417	0,008403361	0,057939645	0,008067039	0,056096133	0,000495187	-0,000475369
5	0,987373737	0,01810036	0,989045455	0,015717187	-0,002409739	0,002413819	0,012658228	0,079794693	0,010954545	0,071339577	0,00263982	-0,002284525
6	0,9784689	0,030725977	0,979851675	0,028773068	-0,001993515	0,001996332	0,021582734	0,119438386	0,020148325	0,113499471	0,002141389	-0,00199907
7	0,793413174	0,264885545	0,798281437	0,259465919	-0,00700196	0,007044923	0,207207207	0,470537297	0,201718563	0,465886014	0,008025204	-0,007812627
8												
9												
10												
SUMA PARCIAL	0,346183862		0,340557613		-0,007254434	0,007324219		0,833766346		0,834351322	0,010671458	-0,005775247

Entropía p = 1,179950208 BITS

Entropía f = 1,174908936 BITS

Entropía relativa D(p/f) = 0,003417023 BITS

Entropía relativa D(f/p) = 0,001548972 BITS

Ojo

DiSCREPANCIA INTRÍNSECA $\delta(p,f)$ = 0,00154897 BITS

DIVERGENCIA JEFFREYS $J(p,f)$ = 0,004966 BITS

El modelo de RL queda a una distancia despreciable de la verdadera distribución

Muerte en UCIP estimada con el índice "PRISM":

Comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística, y una red neuronal artificial. Una propuesta bayesiana

B) Usando como estimador puntual por estrato el ESTIMADOR INTRÍNSECO

"Estratos" de Riesgo Predicho	Nº niños	Exitus (en ese ingreso) = Vivo				Exitus (en ese ingreso) = Muerto			
		Nº Observado Vivos	p(vivir)	Nº Esperado Vivos	f(vivir)	Nº Observado Muertos	p(morir)	Nº Esperado Muertos	f(morir)
1	964	963	0,99861783	961,654	0,997222529	1	0,00138217	2,346	0,002777471
2	120	120	0,997237569	119,62	0,994088398	0	0,002762431	0,38	0,005911602
3	175	174	0,992409867	174,238	0,993764706	1	0,007590133	0,762	0,006235294
4	178	177	0,992537313	177,056	0,992850746	1	0,007462687	0,944	0,007149254
5	197	195	0,988195616	195,331	0,989870152	2	0,011804384	1,669	0,010129848
6	208	204	0,979233227	204,289	0,980618211	4	0,020766773	3,711	0,019381789
7	166	132	0,794	132,813	0,798878	34	0,206	33,187	0,201122
8									
9									
10									
Total	2008	1965		1965,001		43		42,999	

"Estratos"	p(v)	p(v)*log2[1/p(v)]	f(v)	f(v)*log2[1/f(v)]	p(v)*log2[p(v)/f(v)]	f(v)*log2[f(v)/p(v)]	p(m)	p(m)*log2[1/p(m)]	f(morir)	f(m)*log2[1/f(m)]	p(m)*log2[p(m)/f(m)]	f(m)*log2[f(m)/p(m)]
1	0,99861783	0,001992671	0,997222529	0,004001473	0,002014401	-0,002011586	0,00138217	0,013129024	0,002777471	0,023586316	-0,001391619	0,002796459
2	0,997237569	0,003979836	0,994088398	0,00850338	0,004550483	-0,004536113	0,002762431	0,023480237	0,005911602	0,043759069	-0,003032074	0,006488638
3	0,992409867	0,010908585	0,993764706	0,008967524	-0,001953286	0,001955953	0,007590133	0,053447128	0,006235294	0,045675566	0,002153074	-0,00176875
4	0,992537313	0,010726108	0,992850746	0,010277235	-0,000452117	0,000452259	0,007462687	0,052732009	0,007149254	0,050959821	0,000461958	-0,000442556
5	0,988195616	0,016929214	0,989870152	0,01454001	-0,002413801	0,002417891	0,011804384	0,075601574	0,010129848	0,067112712	0,002605349	-0,002235761
6	0,979233227	0,029646857	0,980618211	0,027689266	-0,001996698	0,001999522	0,020766773	0,116077522	0,019381789	0,110265992	0,002067858	-0,001929948
7	0,794	0,264234535	0,798878	0,258798841	-0,007015937	0,00705904	0,206	0,469532454	0,201122	0,465367586	0,007122128	-0,006953479
8												
9												
10												
SUMA PARCIAL	0,338417805		0,332777731	-0,007266955	0,007336966		0,803999949		0,806727063	0,009986674	-0,004045397	

Entropía p = 1,142417755 BITS

Entropía f = 1,139504793 BITS

Entropía relativa D(p/f) = 0,002719719 BITS

Entropía relativa D(f/p) = 0,003291569 BITS



DiSCREPANCIA INTRÍNSECA $\delta(p,f)$ = 0,00271972 BITS
DIVERGENCIA JEFFREYS $J(p,f)$ = 0,00601129 BITS

El modelo de RL queda a una distancia despreciable de la verdadera distribución.

1.2.- MODELO DE RED NEURONAL ARTIFICIAL:

La red se entrenó con la misma muestra de datos de un total de 2032 niños, entre los que se observaron un total de 43 muertos. El modelo final de se ajustó sólo con los 2008 niños de los que se dispuso de todos los datos, con los siguientes parámetros:

nh =

5

rango =

2

nuh =

2.1414e-004

nus =

2.1414e-004

epoch =

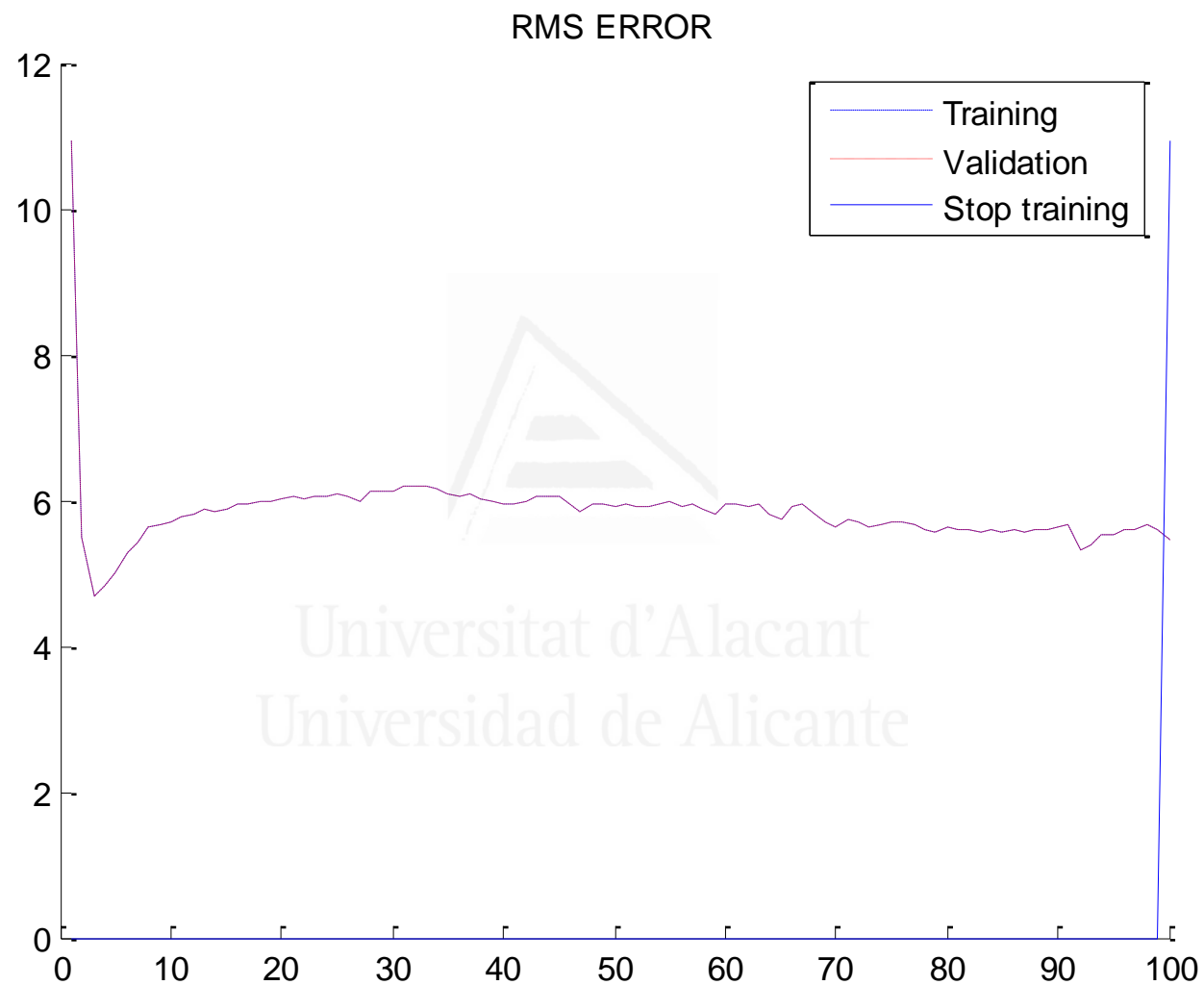
100

Comienza entrenamiento

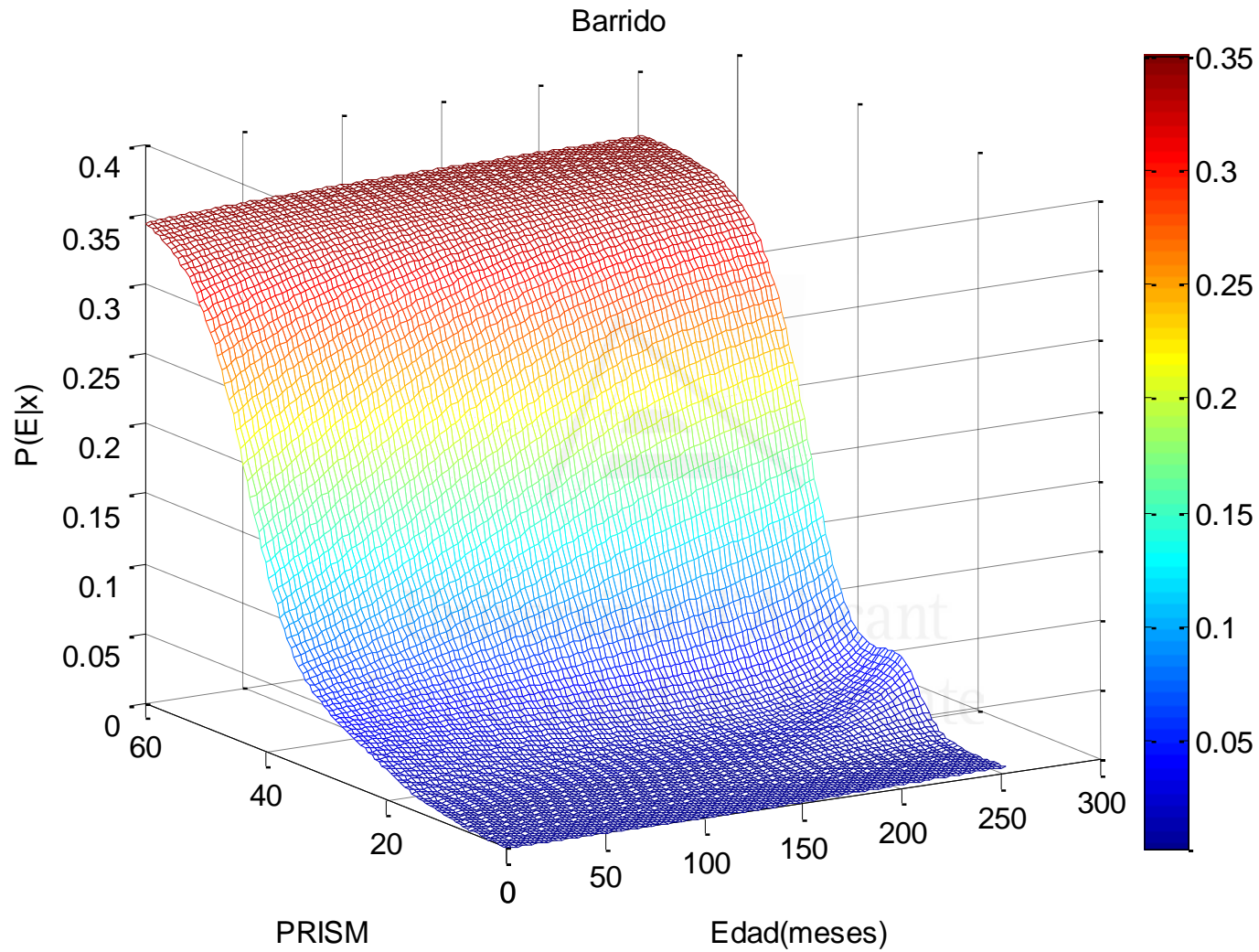
Finaliza entrenamiento

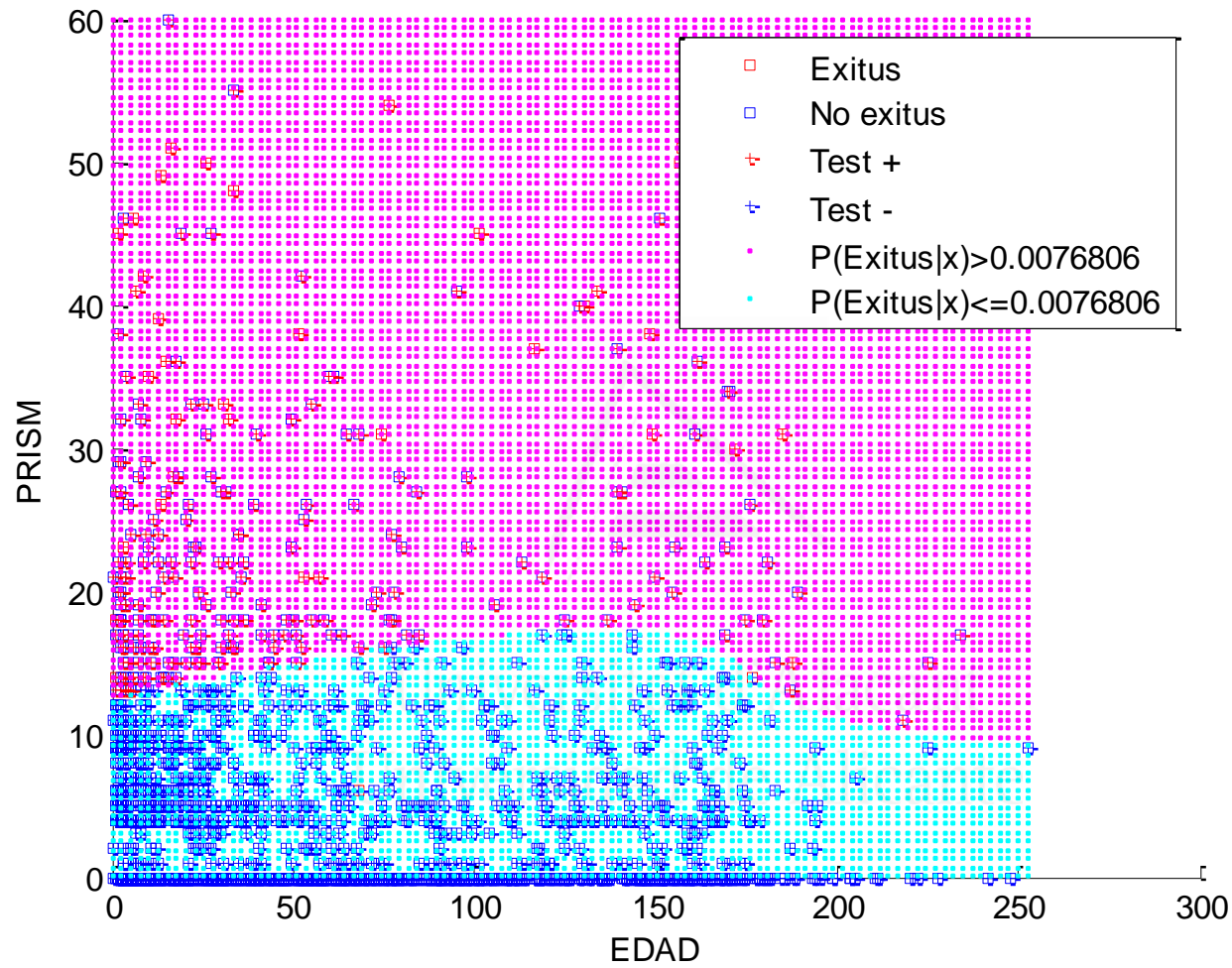
En los siguientes gráficos se representa la probabilidad de éxito estimada por la red en función de los datos de PRISM y EDAD ($p(E/x)$), y la superficie óptima de clasificación, que la red determina como $P(E/x) \geq 0.00768706$.

Muerte en UCIP estimada con el índice "PRISM":
Comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana



Universitat d'Alacant
Universidad de Alicante

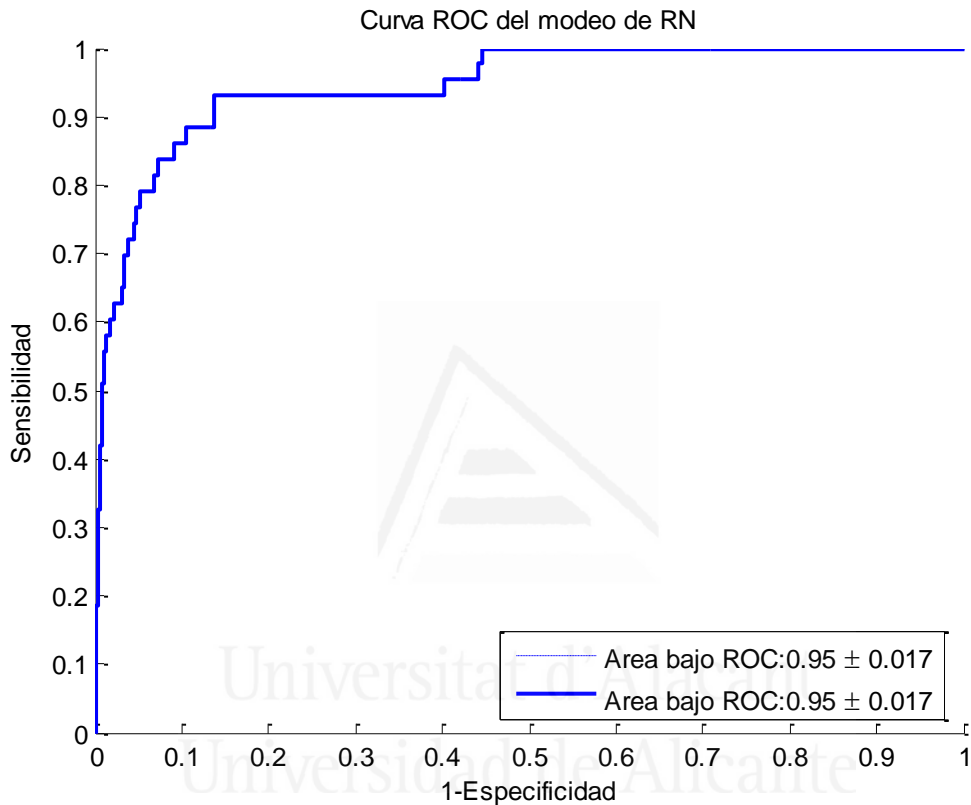




1.2.1.- Evaluación clásica de la exactitud diagnóstica:

1.2.1.1.- Capacidad discriminante:

a) Estudio de la Curva ROC:



La red encontró que el punto de corte óptimo para el máximo rendimiento diagnóstico era una probabilidad pronosticada final $\geq 0,0076806$, que dá al modelo una Sensibilidad de 0,93 y una Especificidad de 0,86.

Área bajo la curva

Área	Error típico ^a	Signif. Asintótica ^b	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
0,95	0,017	0,000	0,917	0,983

a Bajo el supuesto no paramétrico

b Hipótesis nula: área verdadera = 0,5

b) Índices de exactitud diagnóstica:

Pronosticado RN	Observado		
	Muerto	Vivo	
Positivo (Muerto)	40	268	308
Negativo (Vivo)	3	1697	1700
	43	1965	2008

		IC 95%
Sensibilidad	93,0%	85,4% a 100,6%
Especificidad	86,4%	84,8% a 87,9%
Valor predictivo positivo	13,0%	9,2% a 16,7%
Valor predictivo negativo	99,8%	99,6% a 100,0%
Proporción de falsos positivos	13,6%	12,1% a 15,2%
Proporción de falsos negativos	7,0%	-0,6% a 14,6%
Acierto	86,5%	85,0% a 88,0%
Odds ratio diagnóstica	85,25	32,63 a 198,0
Índice J de Youden	0,8	
Razón de Verosimilitudes + [LR(+)]	6,82	5,94 a 7,83
Razón de Verosimilitudes - [LR(-)]	0,08	0,03 a 0,24
Probabilidad pre-prueba (Prevalencia)	2,1%	

1.2.1.2.- Calibración del modelo:

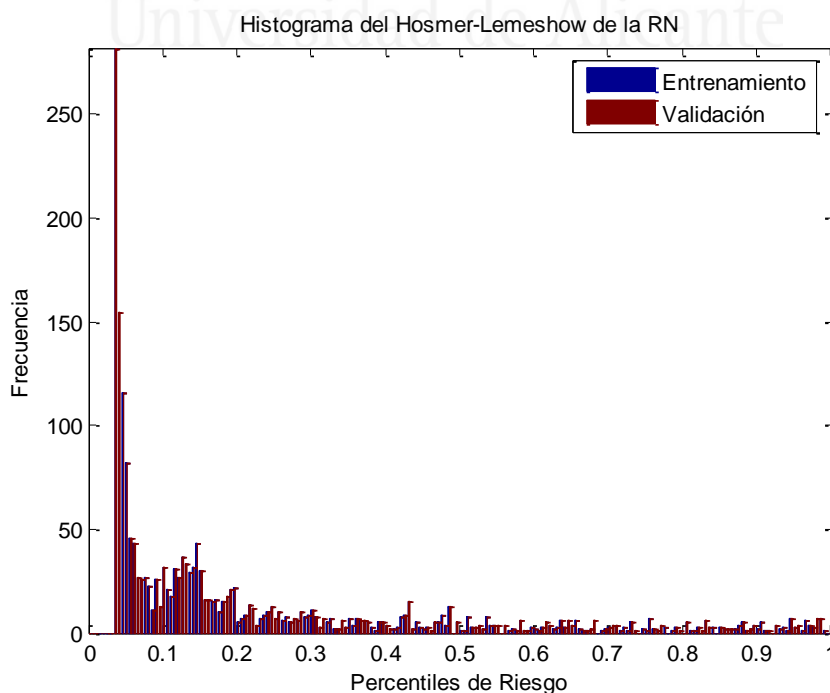


Tabla de contingencias para la prueba de Hosmer y Lemeshow

Estratos de Riesgo	Exitus (en ese ingreso)				Total de niños
	Vivo		Muerto		
	Observado	Esperado	Observado	Esperado	
1	962	961.654	2	2.346	964
2	120	119.62	0	0.38	120
3	174	174.238	1	0.762	175
4	177	177.056	1	0.944	178
5	195	195.331	2	1.669	197
6	204	204.289	4	3.711	208
7	132	132.113	34	33.987	166

Prueba de Hosmer y Lemeshow

Ji-Cuadrado	Grados Libertad	Significación
0,59958	5	0.9881

No se puede descartar la hipótesis nula de "bondad de ajuste"

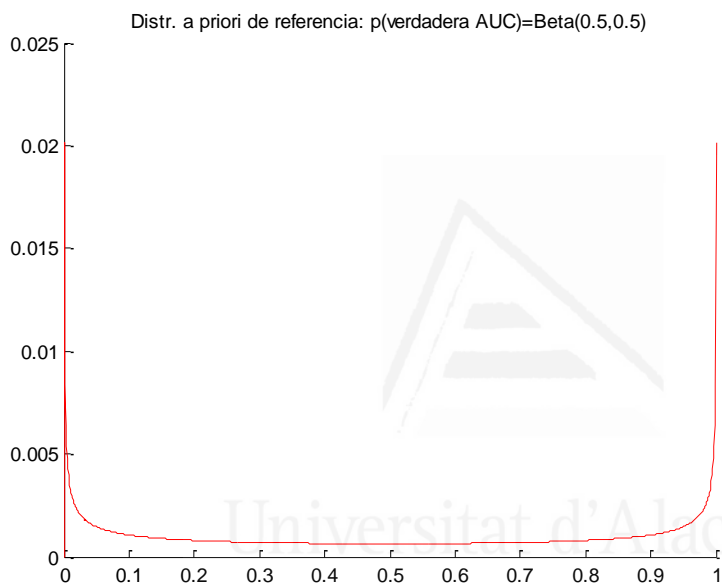
1.2.2.- Evaluación bayesiana de la exactitud diagnóstica:

1.2.2.1.- Capacidad discriminante:

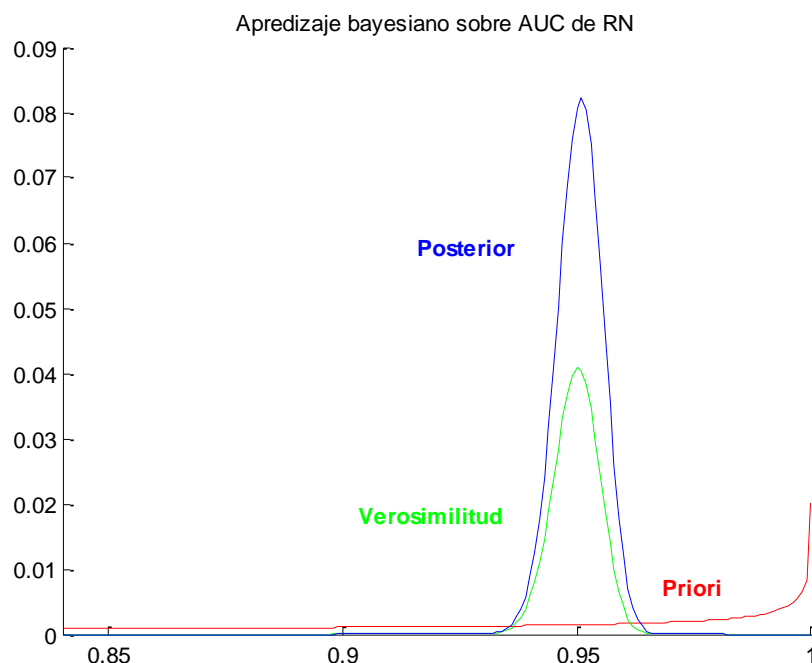
a) Estudio bayesiano de la Curva ROC:

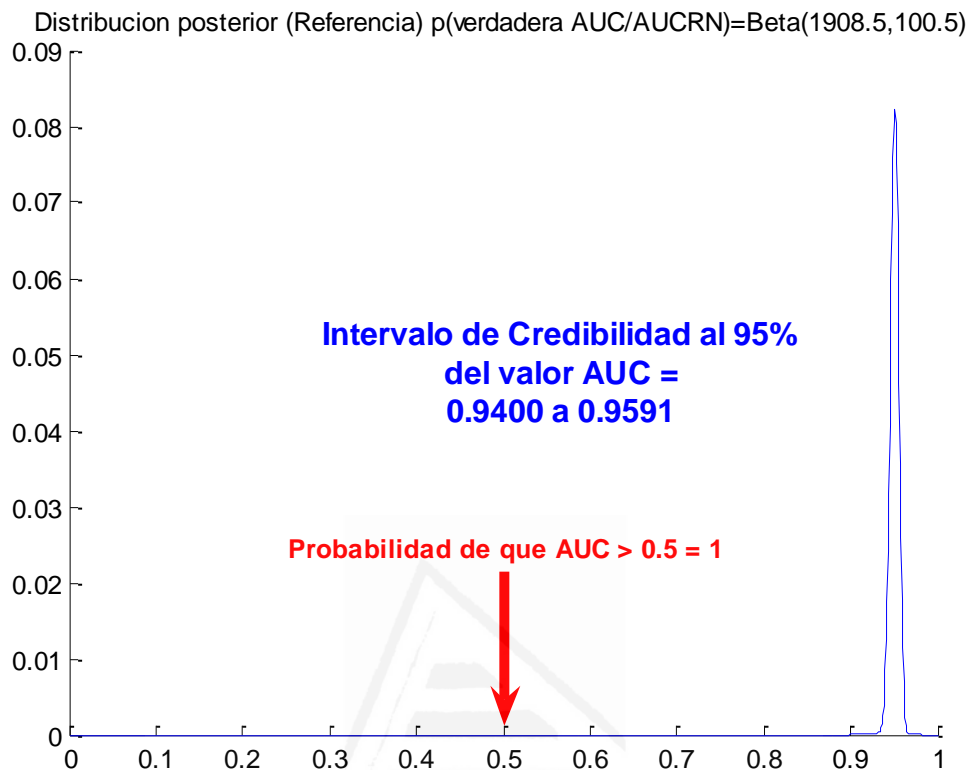
Se considera que la AUC es una probabilidad, por lo que se modeliza con la distribución BETA. Para el análisis bayesiano de referencia, se toma como distribución a priori sobre el parámetro "verdadero valor de la AUC" la **Beta (1/2, 1/2)**. Por el estudio clásico y mediante el método trapezoidal, se sabe que el área bajo curva ROC es 0.95, y que el tamaño muestral es $n = 2008$, por lo que la función de verosimilitud conjugada de la Beta que modeliza el experimento es la Binomial ($r=1908, n=2008, AUC$).

Aplicando el teorema de Bayes, se obtiene que la distribución posterior sobre el parámetro "verdadero valor de la AUC" es la distribución **Beta (1908.5, 100.5)**, cuya media es **AUC = 0.95**. Integrándola, podemos calcular que tenemos una probabilidad del 95% de que el "verdadero valor de la AUC" se sitúe entre 0.9400 y 0.9591 (Intervalo de Credibilidad al 95%). Y además, que la probabilidad de que el "verdadero valor de la AUC" sea mayor que 0.5 -esto es, que el modelo de Red Neuronal clasifique mejor que el azar- es, después de aprender de los datos del experimento, prácticamente 1.



Universitat d'Alacant
Universidad de Alicante





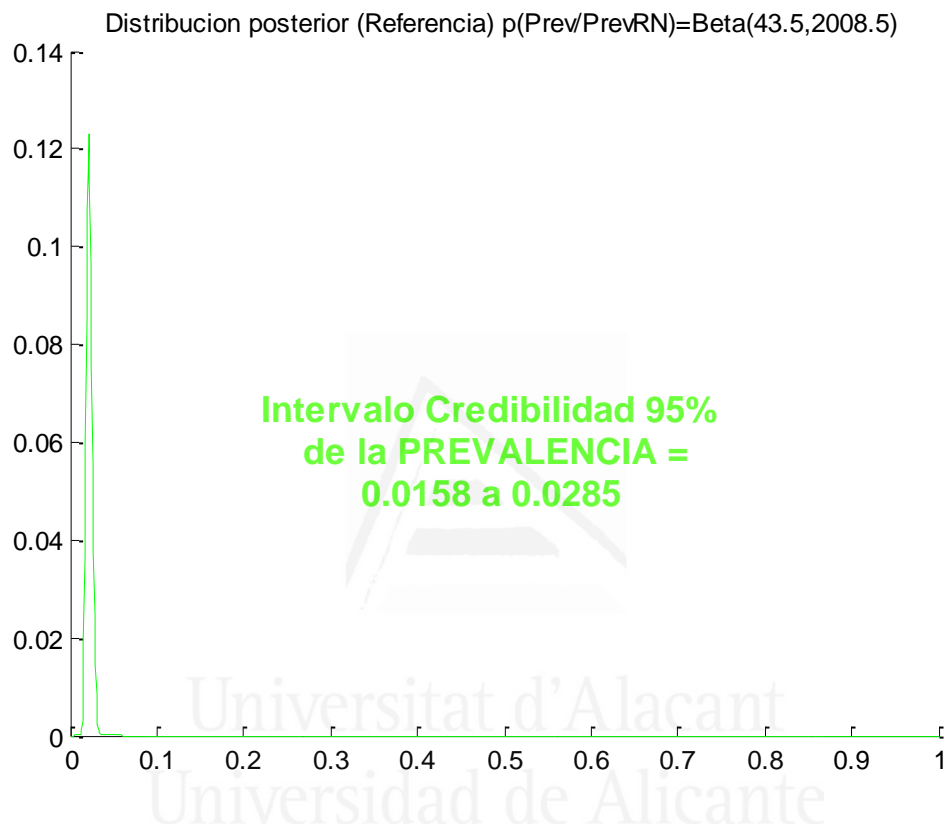
b) Índices bayesianos de exactitud diagnóstica:

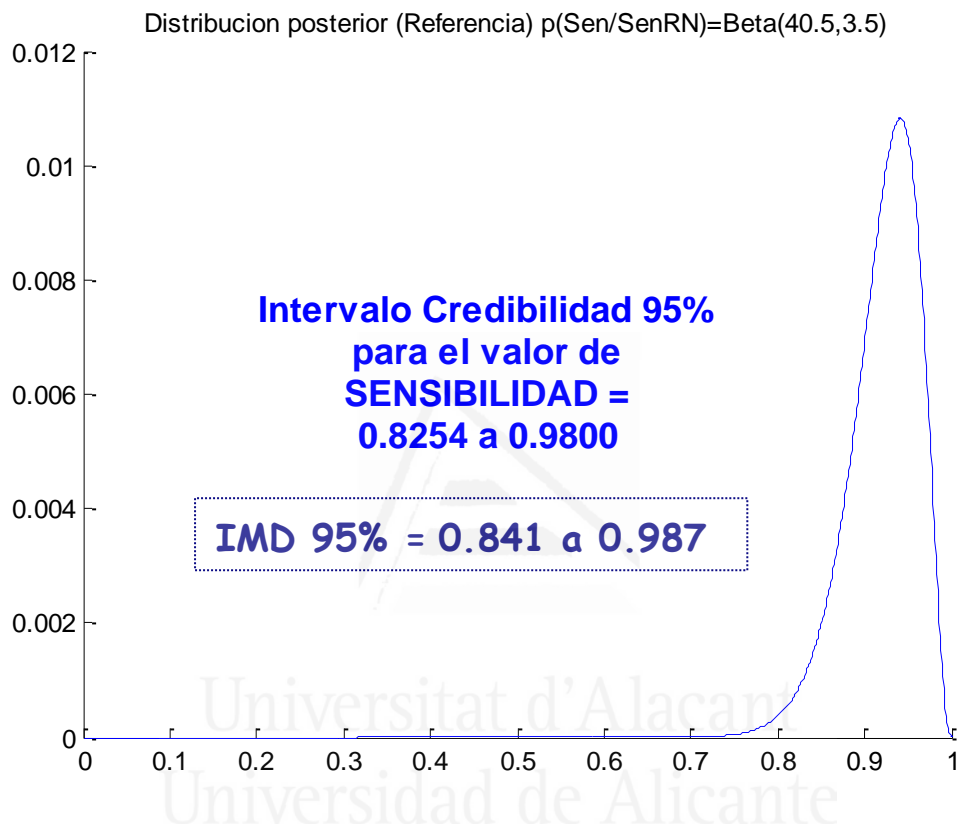
Tanto la Prevalencia como la Sensibilidad, la Especificidad y los Valores Predictivos son probabilidades, por lo que también se modelizan con la distribución BETA. Las Razones de Verosimilitud se modelizan con la distribución NORMAL, usando la transformación logarítmica. Para el análisis bayesiano de referencia, se toma como distribución a priori la **Beta (1/2, 1/2)**. En la tabla siguiente se detallan sus distribuciones de verosimilitud, sus distribuciones posteriores, y los Intervalos de Credibilidad y de Máxima Densidad al 95%. Como en los casos anteriores, los resultados se han obtenido aplicando el Teorema de Bayes.

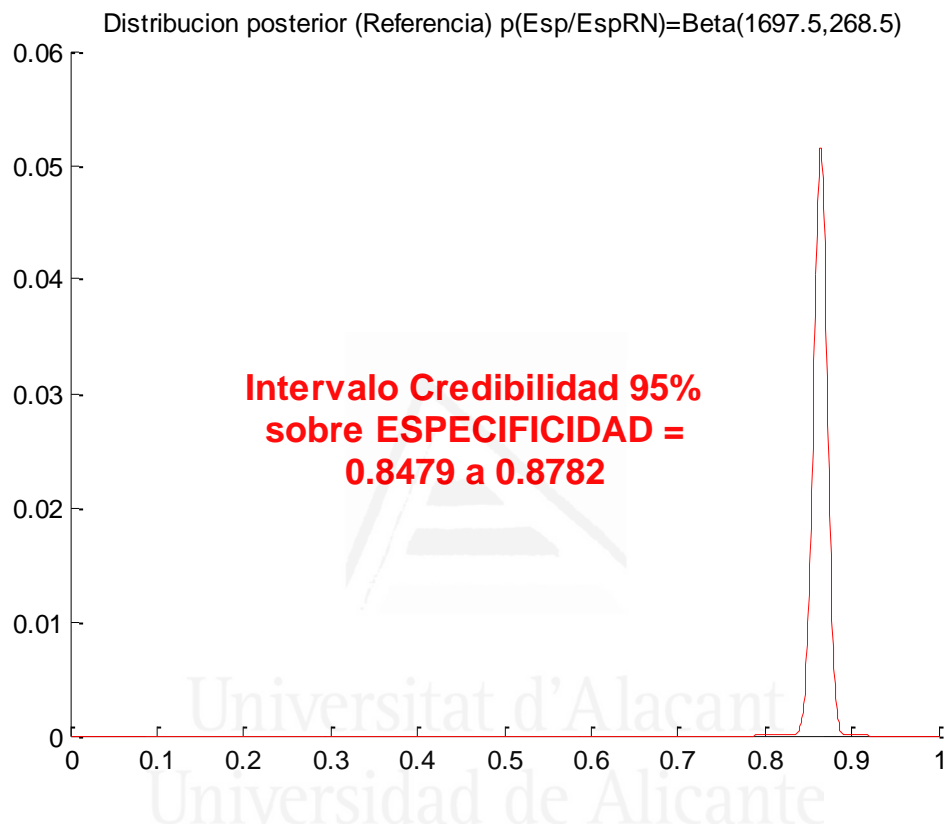
Pronosticado RN	Observado		
	Muerto	Vivo	
Positivo (Muerto)	40	268	308
Negativo (Vivo)	3	1697	1700
	43	1965	2008

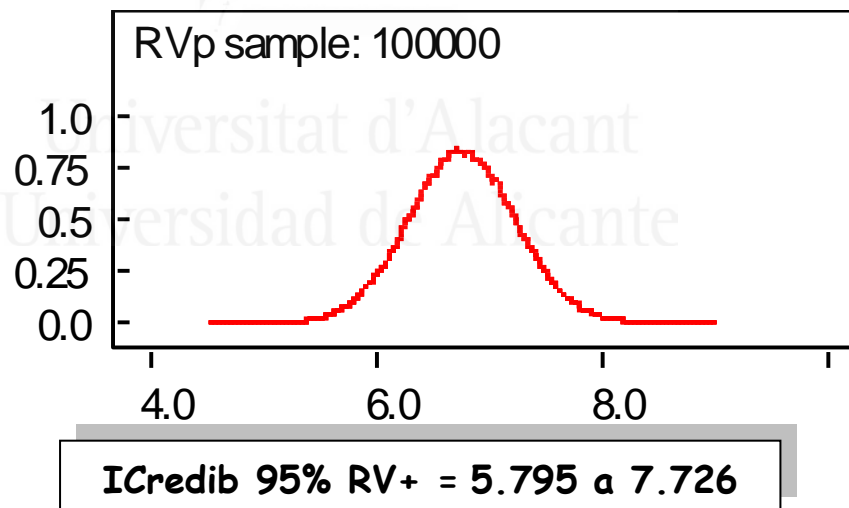
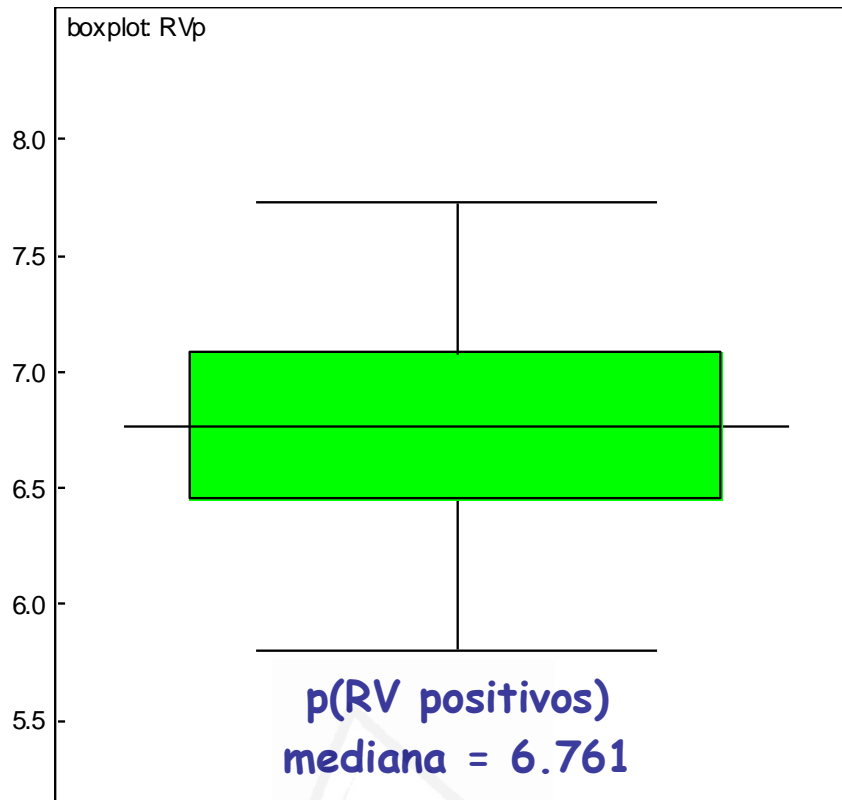
	Priori	Verosimilitud	Posteriori
Prevalencia	Be(1/2,1/2)	Bi(r=43,n=2008,P)	Be(43.5, 1965.5)
Sensibilidad		Bi(r=40,n=43,S)	Be(40.5, 3.5)
Especificidad		Bi(r=1697,n=1965,E)	Be(1697.5, 268.5)
VP Positivos		Bi(r=40,n=308,VPP)	Be(40.5, 268.5)
VP Negativos		Bi(r=1697,n=1700,VPN)	Be(1697.5, 3.5)
Raz Veros +	RVpos = Sens/(1 - Esp)		
Raz Veros -	RVneg = (1 - Sens)/Esp		

	Media (Post)	Int Credib 95%	IMD 95%
Prevalencia	2.165 %	1.58 a 2.85 %	1.58 a 2.85 %
Sensibilidad	92.7 %	82.54 a 98.00 %	84.1 a 98.7 %
Especificidad	86.4 %	84.8 a 87.8 %	84.8 a 87.8 %
VP Positivos	13.00 %	9.6 a 17.1 %	9.4 a 16.9 %
VP Negativos	99.8 %	99.5 a 99.99 %	99.5 a 99.99 %
Raz Veros +	6.761	5.795 a 7.726	
Raz Veros -	0.093	0.023 a 0.206	



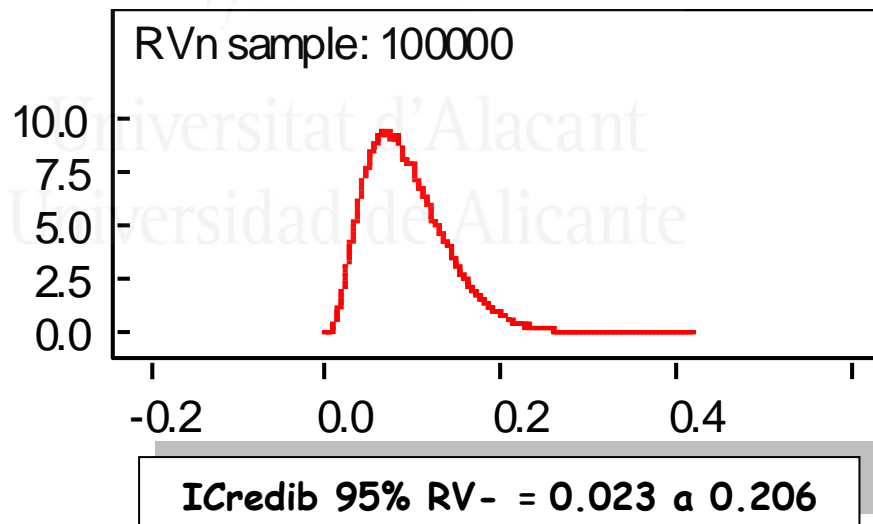
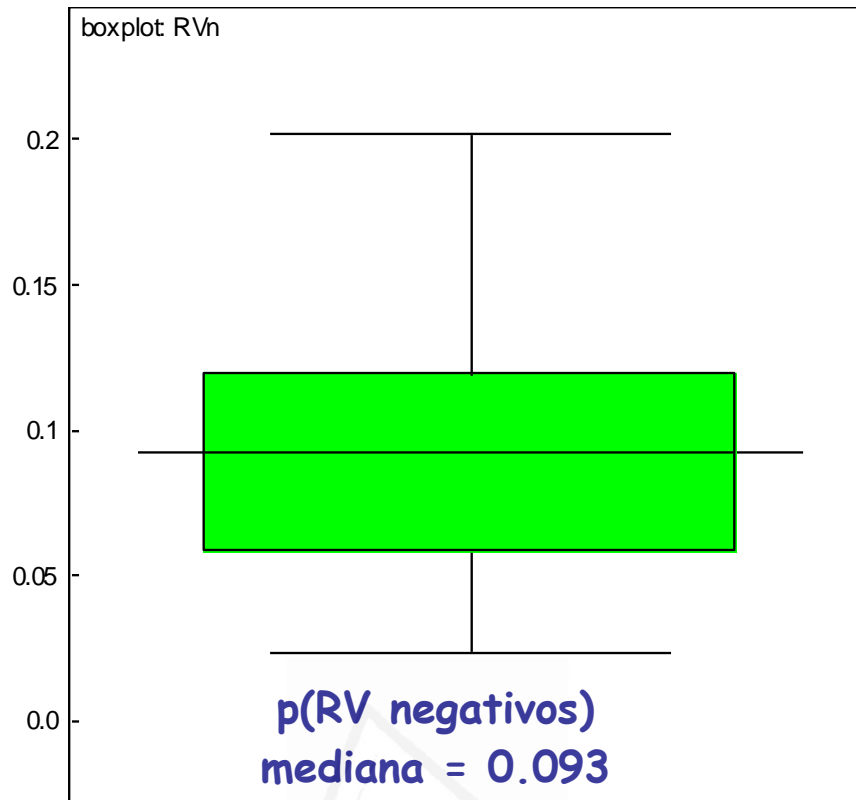






Si consideramos que una RV+ es fuerte cuando es superior a 5 (TABLA 3 y TABLA 4, pag.43), puede calcularse también la probabilidad de que la Red Neuronal presente una $RV+ > 5$ y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV+ > 5) = 0.9998$$



Si consideramos que una RV- es fuerte cuando es inferior a 1/5 (TABLA 3 y TABLA 4, pag.43), puede calcularse la probabilidad de que la Red Neuronal presente una $RV- < 1/5$. y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV- < 1/5) = 0.9735$$

1.2.2.2.- Calibración del modelo:



Universitat d'Alacant
Universidad de Alicante

Muerte en UCIP estimada con el índice "PRISM":

Comparación de la exactitud diagnóstica de las predicciones realizada con un modelo de regresión logística, y una red neuronal artificial. Una propuesta bayesiana

A) Usando como estimador puntual por estrato el ESTIMADOR BAYESIANO CONVENCIONAL

		Exitus (en ese ingreso) = Vivo				Exitus (en ese ingreso) = Muerto			
"Estratos" de Riesgo Predicho	Nº niños	Nº Observado Vivos	p(vivir)	Nº Esperado Vivos	f(vivir)	Nº Observado Muertos	p(morir)	Nº Esperado Muertos	f(morir)
1	964	962	0,997409326	961,654	0,997050777	2	0,002590674	2,346	0,002949223
2	120	120	0,995867769	119,62	0,992727273	0	0,004149378	0,38	0,007272727
3	175	174	0,991477273	174,238	0,992829545	1	0,008547009	0,762	0,007170455
4	178	177	0,991620112	177,056	0,991932961	1	0,008403361	0,944	0,008067039
5	197	195	0,987373737	195,331	0,989045455	2	0,012658228	1,669	0,010954545
6	208	204	0,9784689	204,289	0,979851675	4	0,021582734	3,711	0,020148325
7	166	132	0,793413174	132,113	0,79408982	34	0,207207207	33,987	0,206508982
8									
9									
10									
Total	2008	1964		1964,301		44		43,799	

"Estratos"	p(v)	p(v)*log2[1/p(v)]	f(v)	f(v)*log2[1/f(v)]	p(v)*log2[p(v)/f(v)]	f(v)*log2[f(v)/p(v)]	p(m)	p(m)*log2[1/p(m)]	f(morir)	f(m)*log2[1/f(m)]	p(m)*log2[p(m)/f(m)]	f(m)*log2[f(m)/p(m)]
1	0,997409326	0,003732706	0,997050777	0,004248549	0,00051737	-0,000517184	0,002590674	0,022260251	0,002949223	0,024789543	-0,000484476	0,000551527
2	0,995867769	0,005949216	0,992727273	0,010454081	0,004537937	-0,004523626	0,004149378	0,032833566	0,007272727	0,051660275	-0,003359342	0,005888011
3	0,991477273	0,01224315	0,992829545	0,010307602	-0,001949588	0,001952247	0,008547009	0,058721066	0,007170455	0,051080308	0,002165427	-0,00181667
4	0,991620112	0,012038826	0,991932961	0,011591207	-0,000451275	0,000451417	0,008403361	0,057939645	0,008067039	0,056096133	0,000495187	-0,000475369
5	0,987373737	0,01810036	0,989045455	0,015717187	-0,002409739	0,002413819	0,012658228	0,079794693	0,010954545	0,071339577	0,00263982	-0,002284525
6	0,9784689	0,030725977	0,979851675	0,028773068	-0,001993515	0,001996332	0,021582734	0,119438386	0,020148325	0,113499471	0,002141389	-0,00199907
7	0,793413174	0,264885545	0,79408982	0,264134836	-0,000975779	0,000976611	0,207207207	0,470537297	0,206508982	0,469957356	0,001009027	-0,001005627
8												
9												
10												
SUMA PARCIAL	0,347675781		0,34522653	-0,002724588	0,002749615		0,841524904		0,838422664	0,004607033	-0,001141723	

Entropía p = 1,189200685 BITS

Entropía f = 1,183649194 BITS

Entropía relativa D(p/f) = 0,001882445 BITS

Entropía relativa D(f/p) = 0,001607892 BITS



DISCREPANCIA INTRÍNSECA $\delta(p,f)$ = 0,00160789 BITS

DIVERGENCIA JEFFREYS $J(p,f)$ = 0,00349034 BITS

El modelo de RN queda a una distancia despreciable de la verdadera distribución

Muerte en UCIP estimada con el "PRISM":
 Comparación de la exactitud diagnóstica de las predicciones realizadas con un modelo de regresión logística,
 y una red neuronal artificial. Una propuesta bayesiana

B) Usando como estimador puntual por estrato el ESTIMADOR INTRÍNSECO

		Exitus (en ese ingreso) = Vivo				Exitus (en ese ingreso) = Muerto			
"Estratos" de Riesgo Predicho	Nº niños	Nº Observado Vivos	p(vivir)	Nº Esperado Vivos	f(vivir)	Nº Observado Muertos	p(morir)	Nº Esperado Muertos	f(morir)
1	964	962	0,997581202	961,654	0,997222529	2	0,002418798	2,346	0,002777471
2	120	120	0,997237569	119,62	0,994088398	0	0,002762431	0,38	0,005911602
3	175	174	0,992409867	174,238	0,993764706	1	0,007590133	0,762	0,006235294
4	178	177	0,992537313	177,056	0,992850746	1	0,007462687	0,944	0,007149254
5	197	195	0,988195616	195,331	0,989870152	2	0,011804384	1,669	0,010129848
6	208	204	0,979233227	204,289	0,980618211	4	0,020766773	3,711	0,019381789
7	166	132	0,794	132,113	0,794678	34	0,206	33,987	0,205922
8									
9									
10									
Total	2008	1964		1964,301		44		43,799	

"Estratos"	p(v)	p(v)*log2[1/p(v)]	f(v)	f(v)*log2[1/f(v)]	p(v)*log2[p(v)/f(v)]	f(v)*log2[f(v)/p(v)]	p(m)	p(m)*log2[1/p(m)]	f(morir)	f(m)*log2[1/f(m)]	p(m)*log2[p(m)/f(m)]	f(m)*log2[f(m)/p(m)]
1	0,997581202	0,003485363	0,997222529	0,004001473	0,000517549	-0,000517363	0,002418798	0,021022965	0,002777471	0,023586316	-0,000482506	0,000554054
2	0,997237569	0,003979836	0,994088398	0,00850338	0,004550483	-0,004536113	0,002762431	0,023480237	0,005911602	0,043759069	-0,003032074	0,006488638
3	0,992409867	0,010908585	0,993764706	0,008967524	-0,001953286	0,001955953	0,007590133	0,053447128	0,006235294	0,045675566	0,002153074	-0,00176875
4	0,992537313	0,010726108	0,992850746	0,010277235	-0,000452117	0,000452259	0,007462687	0,052732009	0,007149254	0,050959821	0,000461958	-0,000442556
5	0,988195616	0,016929214	0,989870152	0,01454001	-0,002413801	0,002417891	0,011804384	0,075601574	0,010129848	0,067112712	0,002605349	-0,002235761
6	0,979233227	0,029646857	0,980618211	0,027689266	-0,001996698	0,001999522	0,020766773	0,116077522	0,019381789	0,110265992	0,002067858	-0,001929948
7	0,794	0,264234535	0,794678	0,263481602	-0,00097773	0,000978565	0,206	0,469532454	0,205922	0,469467179	0,000112552	-0,000112509
8												
9												
10												
SUMA PARCIAL	0,339910498		0,337460492	-0,0027256	0,002750714		0,811893889		0,810826655	0,003886211	0,000553168	

Entropía p = 1,151804387 BITS

Entropía f = 1,148287147 BITS

Entropía relativa D(p/f) = 0,001160611 BITS

Entropía relativa D(f/p) = 0,003303882 BITS



DISCREPANCIA INTRÍNSECA $\delta(p,f)$ = 0,00116061 BITS
DIVERGENCIA JEFFREYS $J(p,f)$ = 0,00446449 BITS

El modelo de RN queda a una distancia despreciable de la verdadera distribución.

2.- 2ª FASE: VALIDACIÓN DE LOS TESTS

2.1.- MODELO PREDICTIVO DE REGRESIÓN LOGÍSTICA:

La muestra de validación estuvo formada por los datos de un total de 194 niños, entre los que se observaron un total de 11 muertos. Para cada uno de los 180 niños de los que se dispuso de todos los datos de PRISM, se calculó una probabilidad de muerte utilizando el modelo logístico final obtenido durante la fase de desarrollo. Así, para cada niño, se calculó:

$$P(\text{muerte} / \text{estimaciónRL}) = \frac{1}{1 + e^{[-(-6.016 + 0.146 \bullet PRISM)]}}$$

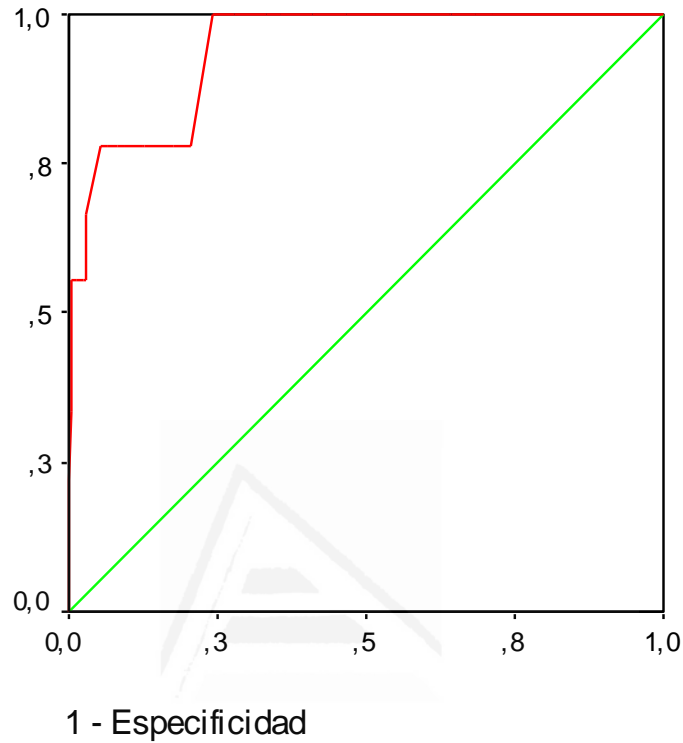
Utilizando como punto de corte $p(\text{muerte}/\text{estimaciónRL}) > 0.011$, esta predicción se comparó con el resultado final de muerte en ese ingreso, evaluando así su rendimiento diagnóstico.

2.1.1.- Evaluación clásica de la exactitud diagnóstica:

2.1.1.1.- Capacidad discriminante:

a) Estudio de la Curva ROC:

Curva ROC



Los segmentos diagonales son producidos por los empates.

Área bajo la curva

Área	Error típico	Signif. Asintótica ^b	Intervalo de confianza asintótico al 95%		Método
			Límite inferior	Límite superior	
0,9412	0,0325	0,000	0,8774	1,0049	DeLong
	0,0549	0,000	0,8336	1,0488	Hanley & McNeil

^b Hipótesis nula: área verdadera = 0,5

b) Índices de exactitud diagnóstica:

Pronosticado RL	Observado		
	Muerto	Vivo	
Positivo (Muerto)	9	41	50
Negativo (Vivo)	0	130	130
	9	171	180

		IC 95%
Sensibilidad	100,0%	100,0% a 100,0%
Especificidad	76,0%	69,6% a 82,4%
Valor predictivo positivo	18,0%	7,4% a 18,6%
Valor predictivo negativo	100,0%	100,0% a 100,0%
Proporción de falsos positivos	24,0%	17,6% a 34,4%
Proporción de falsos negativos	0,0%	0,0% a 0,0%
Acierto	77,2%	71,1% a 83,3%
Odds ratio diagnóstica	36,85	14,40 a 94,31
Índice J de Youden	0,8	
Razón de Verosimilitudes + [LR(+)]	4,17	3,19 a 5,45
Razón de Verosimilitudes - [LR(-)]	0,00	No se puede calcular
Probabilidad pre-prueba (Prevalencia)	5,0%	

2.1.1.2.- Calibración del modelo:

Tabla de contingencias para la prueba de Hosmer y Lemeshow

Estratos de Riesgo	Exitus (en ese ingreso)				Total de niños
	Vivo		Muerto		
	Observado	Esperado	Observado	Esperado	
1	49	48.8808	0	0.1192	49
2	5	4.9859	0	0.0141	5
3	29	28.8820	0	0.1180	29
4	8	7.9597	0	0.0403	8
5	19	18.8812	0	0.1188	19
6	20	19.8151	0	0.1849	20
7	19	18.7545	2	2.2455	21
8	11	10.7406	0	0.2594	11
9	11	12.1766	7	5.8234	18

Prueba de Hosmer y Lemeshow

Ji-Cuadrado	Grados Libertad	Significación
1.2472	7	0.9898

No se puede descartar la hipótesis nula de "bondad de ajuste"

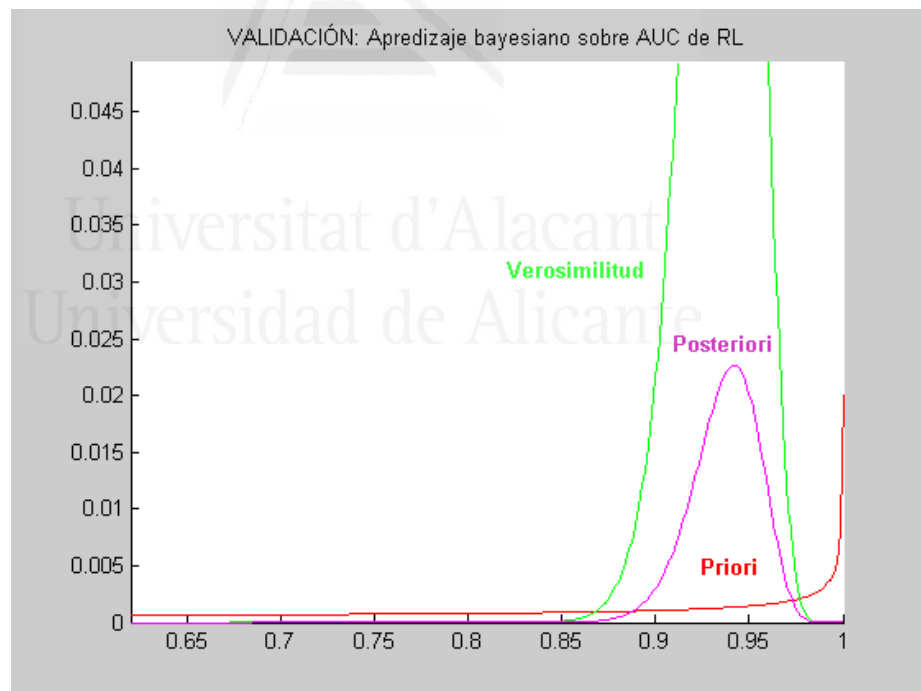
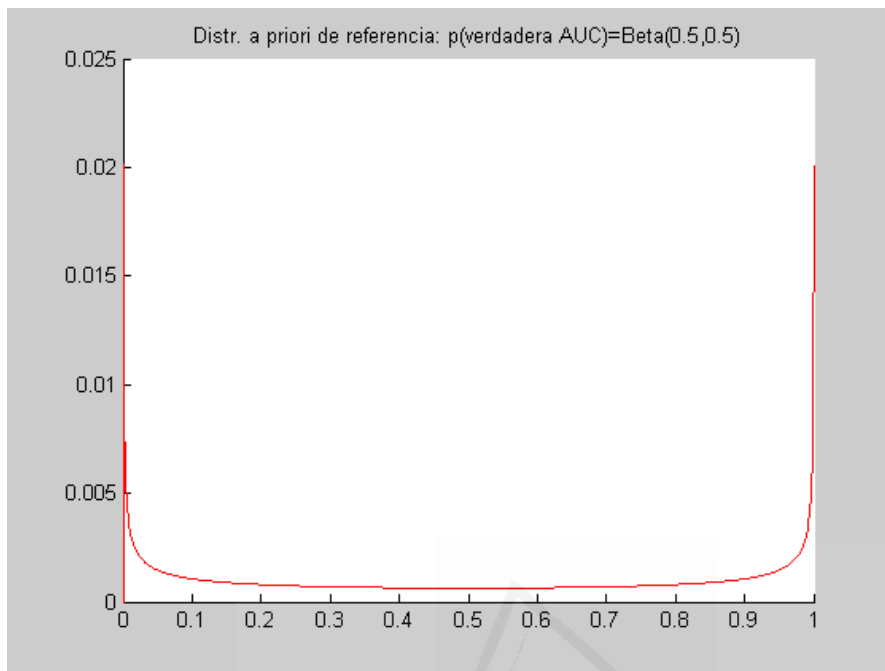
2.1.2.- Evaluación bayesiana de la exactitud diagnóstica:

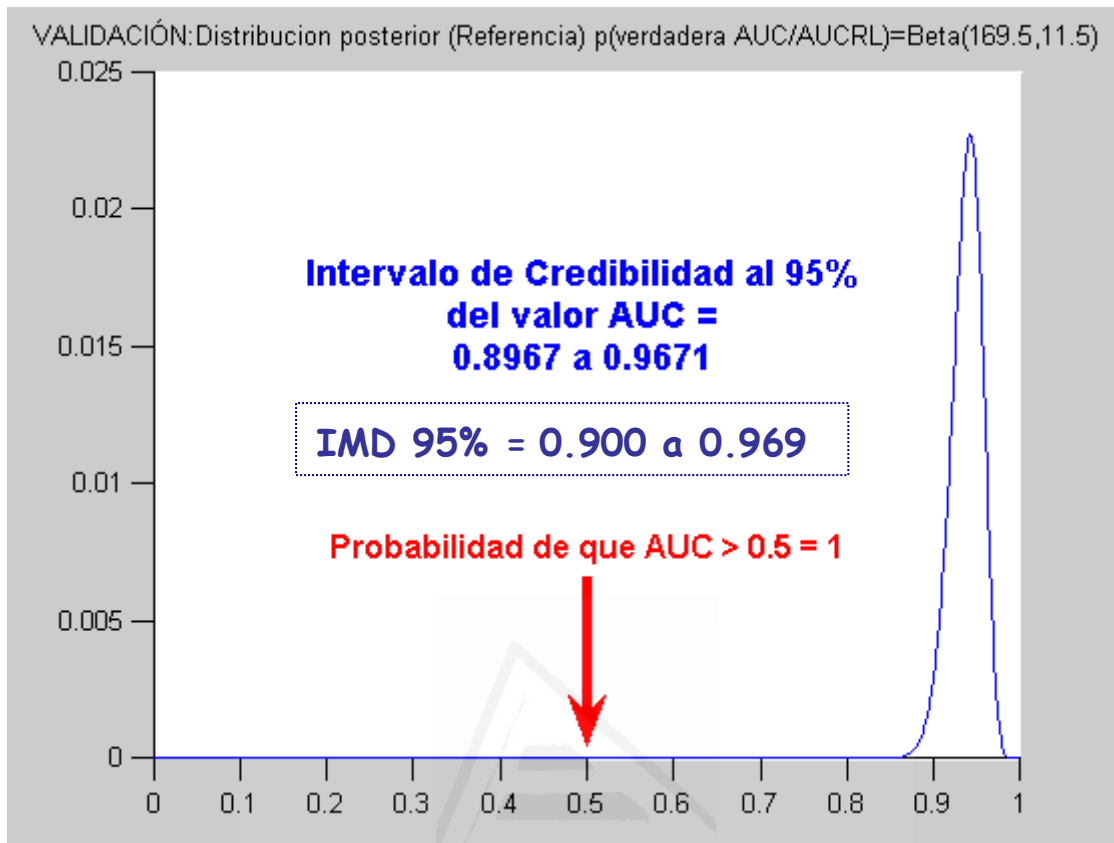
2.1.2.1.- Capacidad discriminante:

a) Estudio bayesiano de la Curva ROC:

Se considera que la AUC es una probabilidad, por lo que se modeliza con la distribución BETA. Para el análisis bayesiano de referencia, se toma como distribución a priori sobre el parámetro "verdadero valor de la AUC" la **Beta (1/2, 1/2)**. Por el estudio clásico y mediante el método trapezoidal, se sabe que el área bajo curva ROC es 0.941, y que el tamaño muestral es $n = 180$, por lo que la función de verosimilitud conjugada de la Beta que modeliza el experimento es la Binomial ($r=169, n=180, AUC$).

Aplicando el teorema de Bayes, se obtiene que la distribución posterior sobre el parámetro "verdadero valor de la AUC" es la distribución **Beta (169.5, 11.5)**, cuya media es **AUC = 0.9365**. Integrándola, podemos calcular que tenemos una probabilidad del 95% de que el "verdadero valor de la AUC" se sitúe entre 0.897 y 0.967 (Intervalo de Credibilidad al 95%). Ya que es asimétrica, convendría estimar el IMD. En este caso $IMD\ 95\% = 0.9$ a 0.969. La probabilidad de que el "verdadero valor de la AUC" sea mayor que 0.5 -esto es, que el modelo de regresión logística clasifique mejor que el azar- es, después de aprender de los datos del experimento, prácticamente 1.





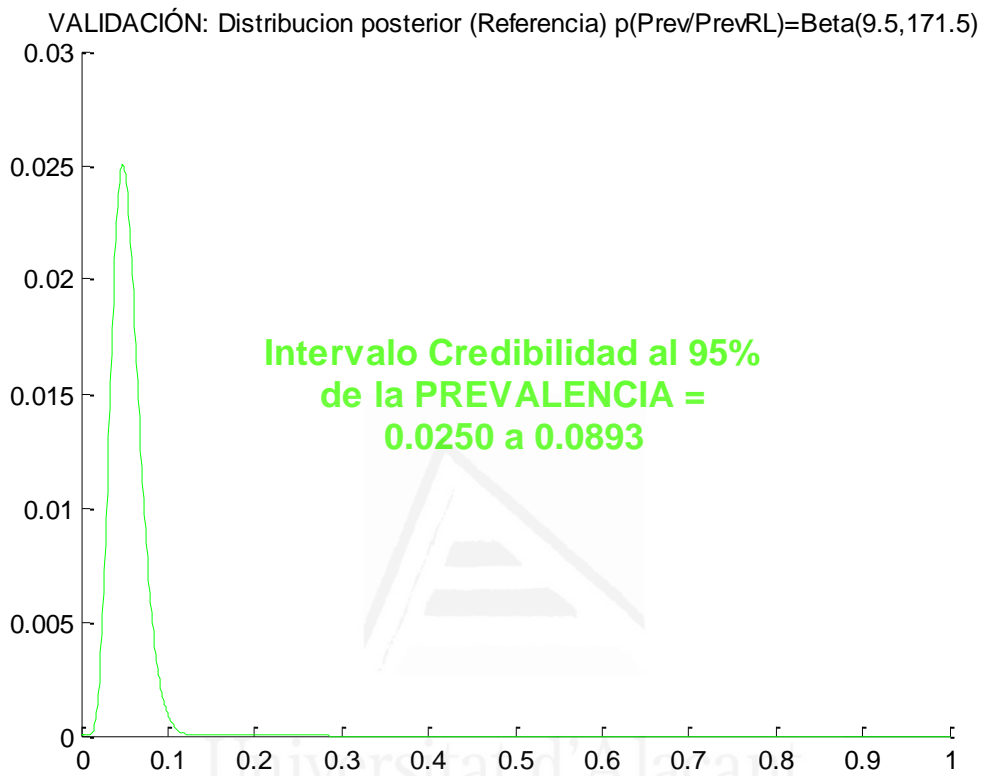
b) Índices bayesianos de exactitud diagnóstica:

Tanto la Prevalencia como la Sensibilidad, la Especificidad y los Valores Predictivos son probabilidades, por lo que también se modelizan con la distribución BETA. Las Razones de Verosimilitud se modelizan con la distribución NORMAL, usando la transformación logarítmica. Para el análisis bayesiano de referencia, se toma como distribución a priori la **Beta (1/2, 1/2)**. En la tabla siguiente se detallan sus distribuciones de verosimilitud, sus distribuciones posteriores, y los Intervalos de Credibilidad y de Máxima Densidad al 95%. Los resultados se han obtenido aplicando el Teorema de Bayes.

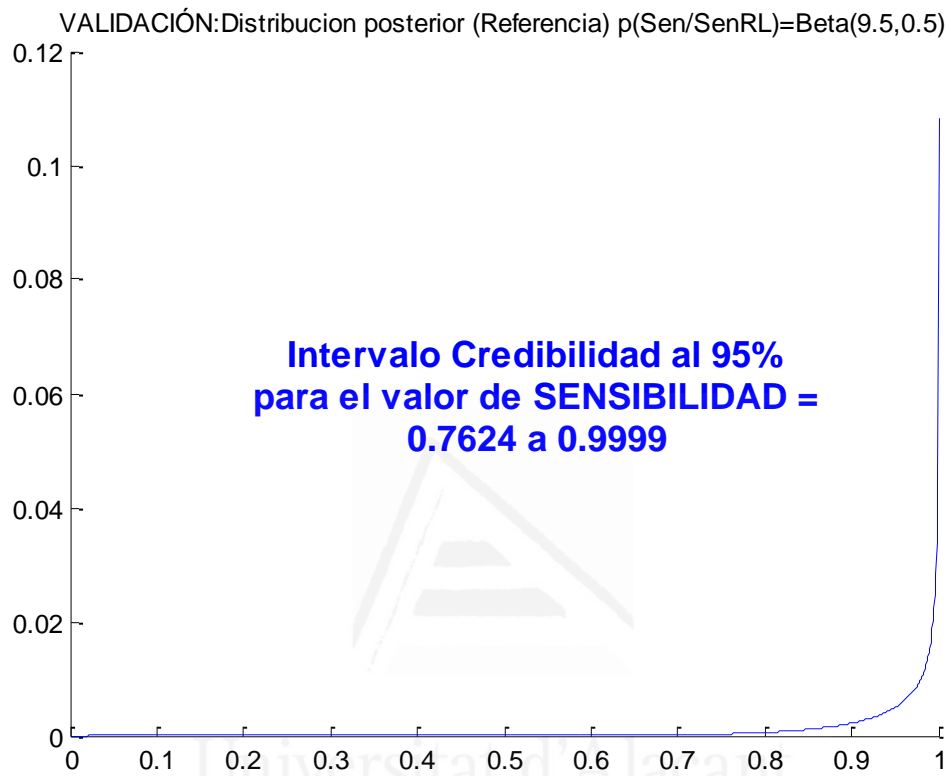
Pronosticado RL	Observado		
	Muerto	Vivo	
Positivo (Muerto)	9	41	50
Negativo (Vivo)	0	130	130
	9	171	180

	Priori	Verosimilitud	Posteriori
Prevalencia	Be(1/2,1/2)	Bi(r=9,n=180,P)	Be(9.5, 171.5)
Sensibilidad		Bi(r=9,n=9,S)	Be(9.5, 0.5)
Especificidad		Bi(r=130,n=171,E)	Be(130.5, 41.5)
VP Positivos		Bi(r=9,n=50,VPP)	Be(9.5, 41.5)
VP Negativos		Bi(r=130,n=130,VPN)	Be(130.5, 0.5)
Raz Veros +	RVpos = Sens/(1 - Esp)		
Raz Veros -	RVneg = (1 - Sens)/Esp		

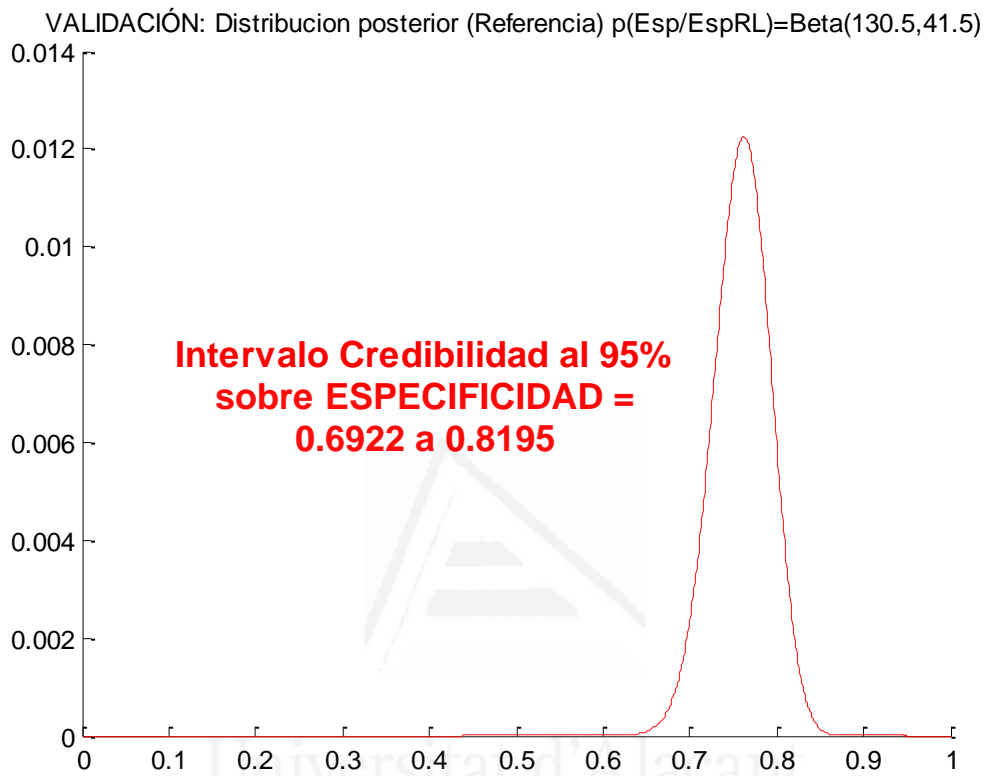
	Media (Post)	Int Credib 95%	IMD 95%
Prevalencia	5.00 %	2.5 a 8.9 %	2.2 a 8.5 %
Sensibilidad	95.0 %	76.24 a 99.999 %	76.24 a 99.999 %
Especificidad	76.0 %	69.22 a 81.95 %	69.4 a 82.1 %
VP Positivos	19.0 %	9.3 a 30.3 %	8.6 a 29.3 %
VP Negativos	99.8 %	98.1 a 99.999 %	98.3 a 99.999 %
Raz Veros +	4.011	2.477 a 5.54	
Raz Veros -	0.03196	0.00007 a 0.3156	



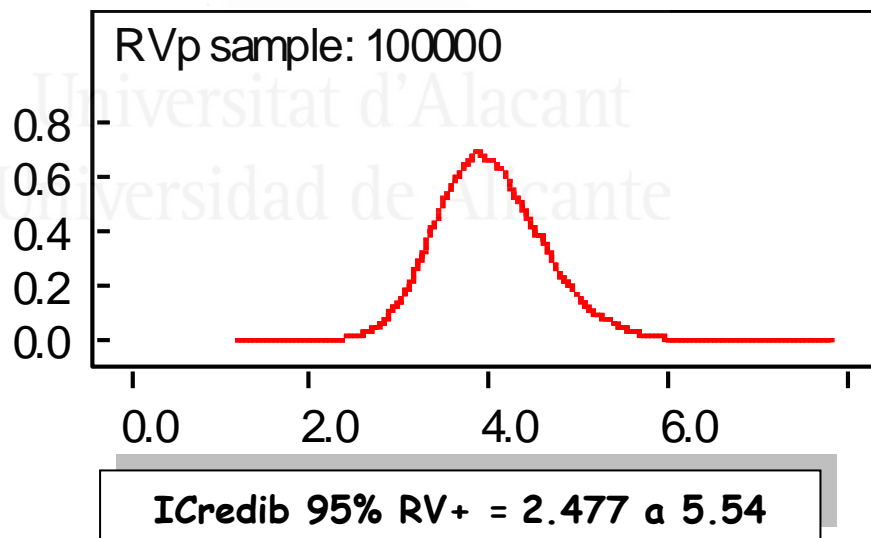
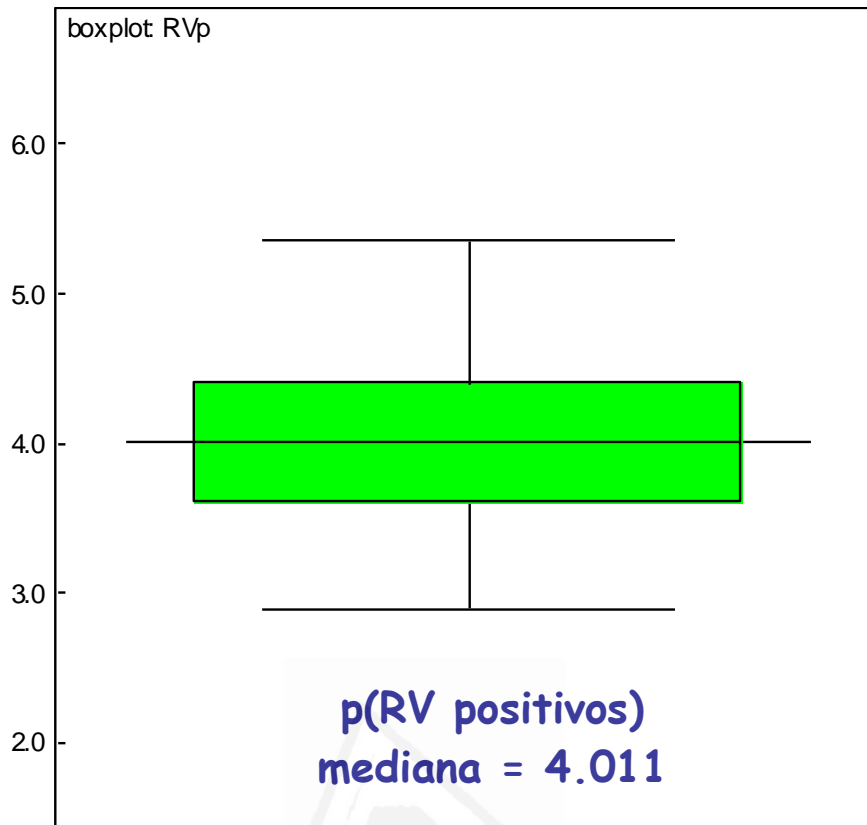
Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

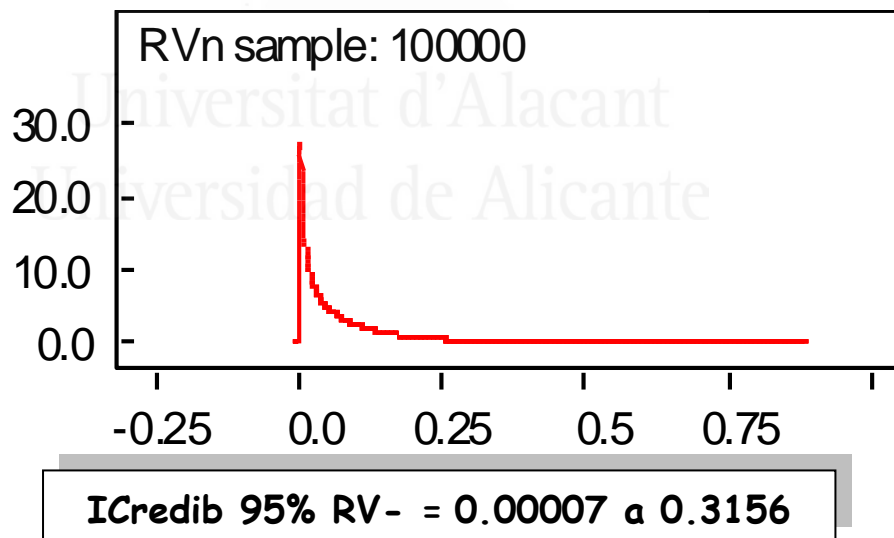
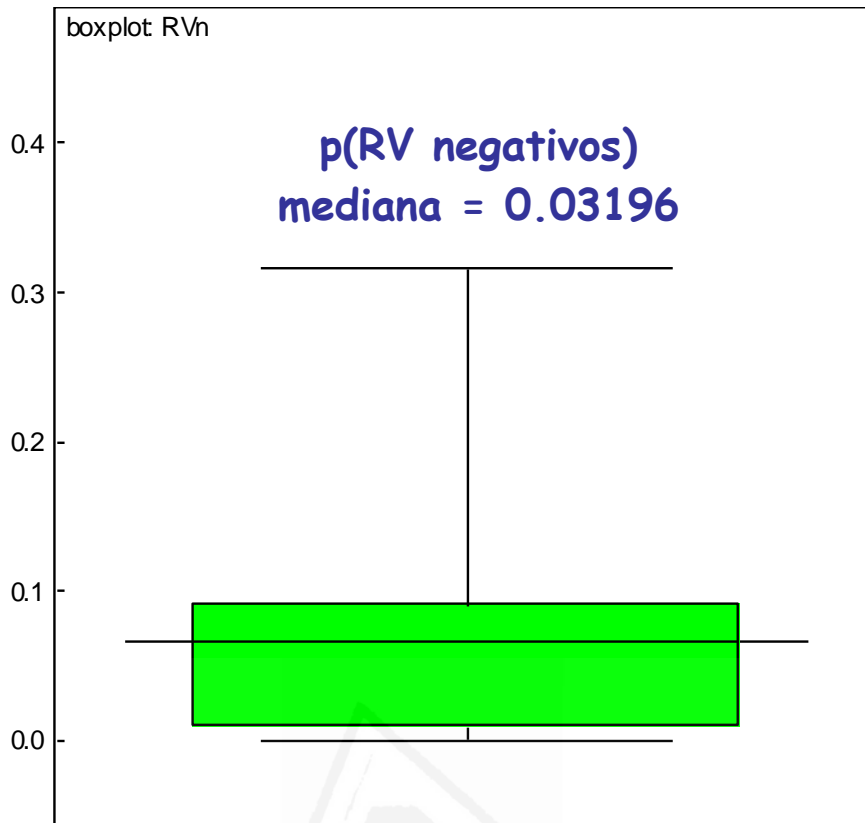


Universitat d'Alacant
Universidad de Alicante



Si consideramos que una RV+ es fuerte cuando es superior a 5 (TABLA 3 y TABLA 4, pag.43), puede calcularse también la probabilidad de que la Regresión Logística presente una $RV+ > 5$ y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV+ > 5) = 0.06358$$



Si consideramos que una RV- es fuerte cuando es inferior a 1/5 (TABLA 3 y TABLA 4, pag.43), puede calcularse la probabilidad de que la Regresión Logística presente una RV- < 1/5. y sea así un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV- < 1/5) = 0.9183$$

2.1.2.2.- Calibración del modelo:

Tabla de contingencias para la prueba de Hosmer y Lemeshow

Estratos de Riesgo	Exitus (en ese ingreso)				Total de niños
	Vivo		Muerto		
	Observado	Esperado	Observado	Esperado	
1	49	48.8808	0	0.1192	49
2	5	4.9859	0	0.0141	5
3	29	28.8820	0	0.1180	29
4	8	7.9597	0	0.0403	8
5	19	18.8812	0	0.1188	19
6	20	19.8151	0	0.1849	20
7	19	18.7545	2	2.2455	21
8	11	10.7406	0	0.2594	11
9	11	12.1766	7	5.8234	18

a) Usando el estimador bayesiano convencional:

Entropía Observados = 2.9859 bits

Entropía Esperados = 3.1980 bits

Entropía Relativa Obs a Esp: $D(o||e) = 0.0235$ bits

Entropía Relativa Esp a Obs: $D(e||o) = 0.0245$ bits

Divergencia de Jeffreys $J(o,e) = 0.048$ bits

Discrepancia Intrínseca $\delta(o,e) = 0.0235$ bits

El modelo de RL queda a una distancia despreciable de la observada

b) Usando el estimador bayesiano intrínseco:

Entropía Observados = 2.5947 bits

Entropía Esperados = 2.8356 bits

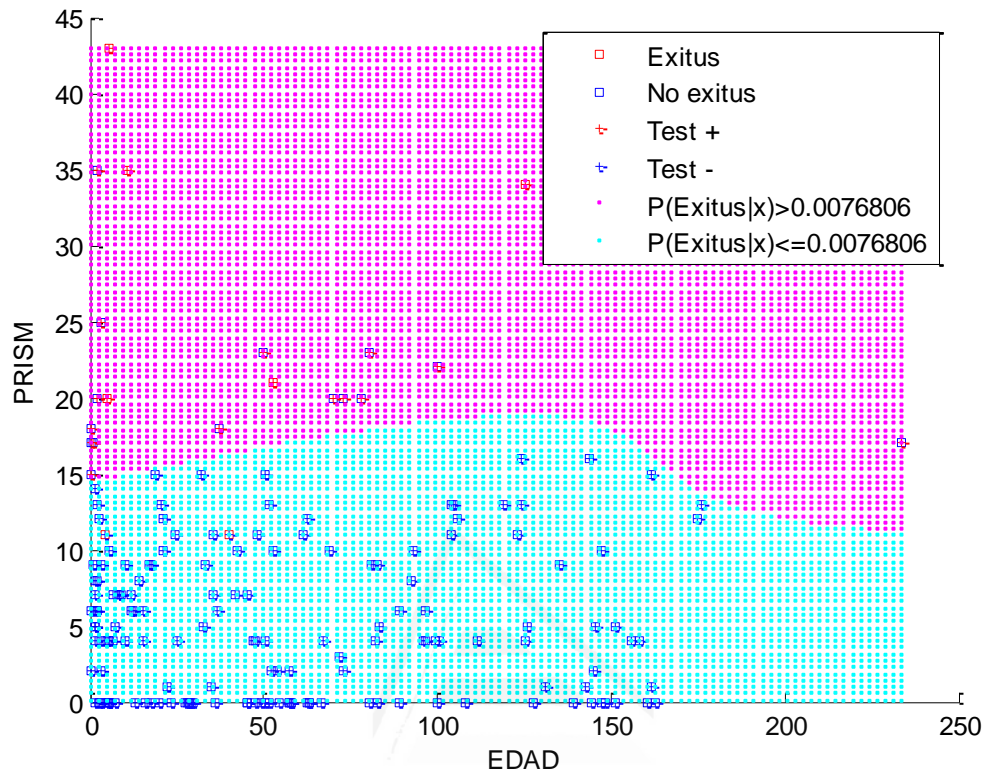
Entropía Relativa Obs a Esp: $D(o||e) = 0.0280$ bitsEntropía Relativa Esp a Obs $D(e||o) = 0.0302$ bits**Divergencia de Jeffreys $J(o,e) = 0.0582$ bits****Discrepancia Intrínseca $\delta(o,e) = 0.0280$ bits**

El modelo de RL queda a una distancia despreciable de la observada

2.2.- MODELO DE RED NEURONAL ARTIFICIAL:

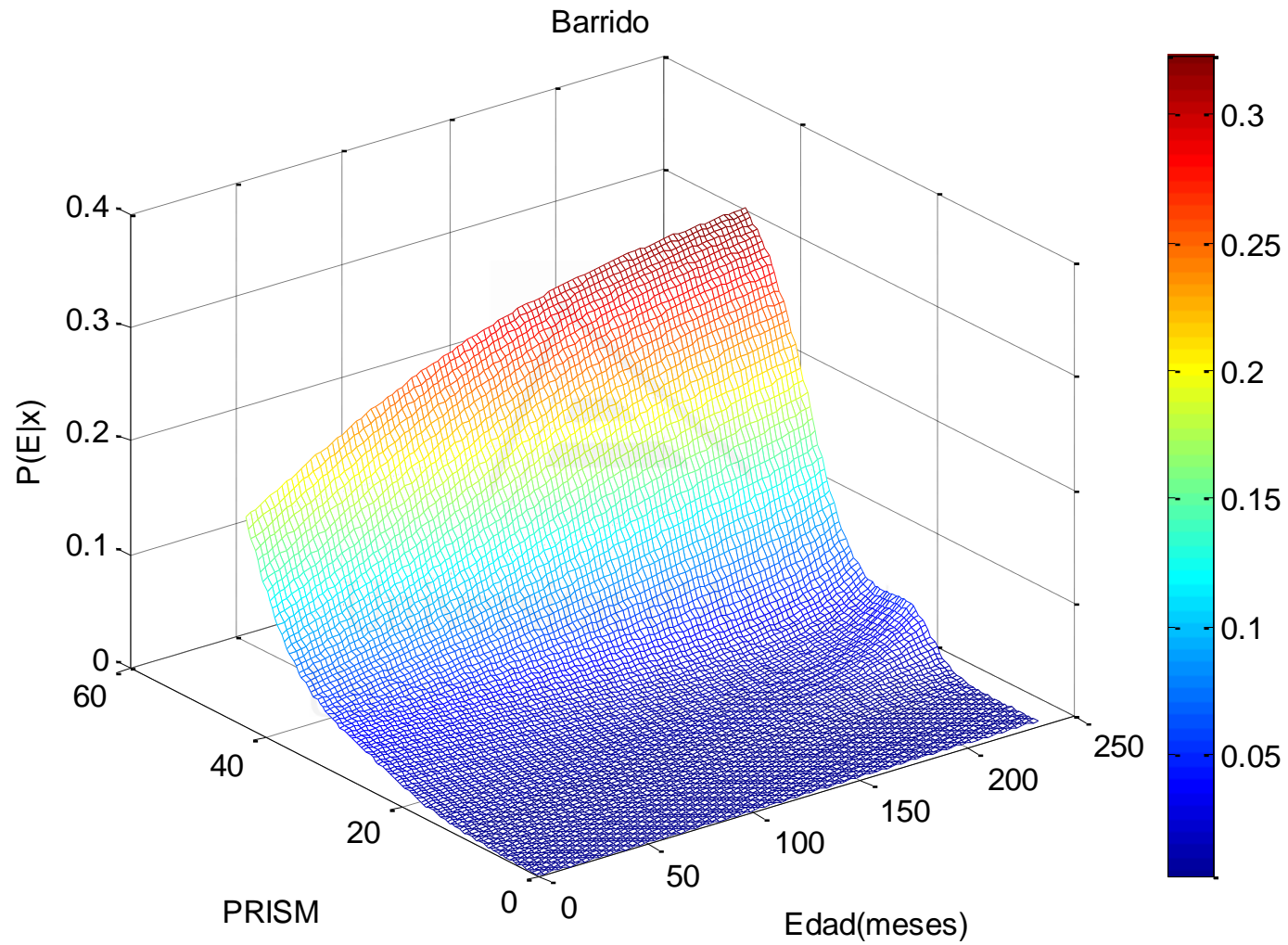
A la misma red que se había entrenado durante la fase de desarrollo, se la enfrentó con los datos de los 180 niños de la cohorte de validación, para evaluar su rendimiento diagnóstico. Se utilizó el mismo punto de corte: una probabilidad pronosticada por la red > 0.0076806 .

En los siguientes gráficos se representa la superficie de clasificación, usando como punto de corte $p(E/x) \geq 0.00768706$, y, para cada niño, la $p(E/x)$: la probabilidad de éxito estimada por la red en función de los datos de PRISM y EDAD.



Universitat d'Alacant
 Universidad de Alicante

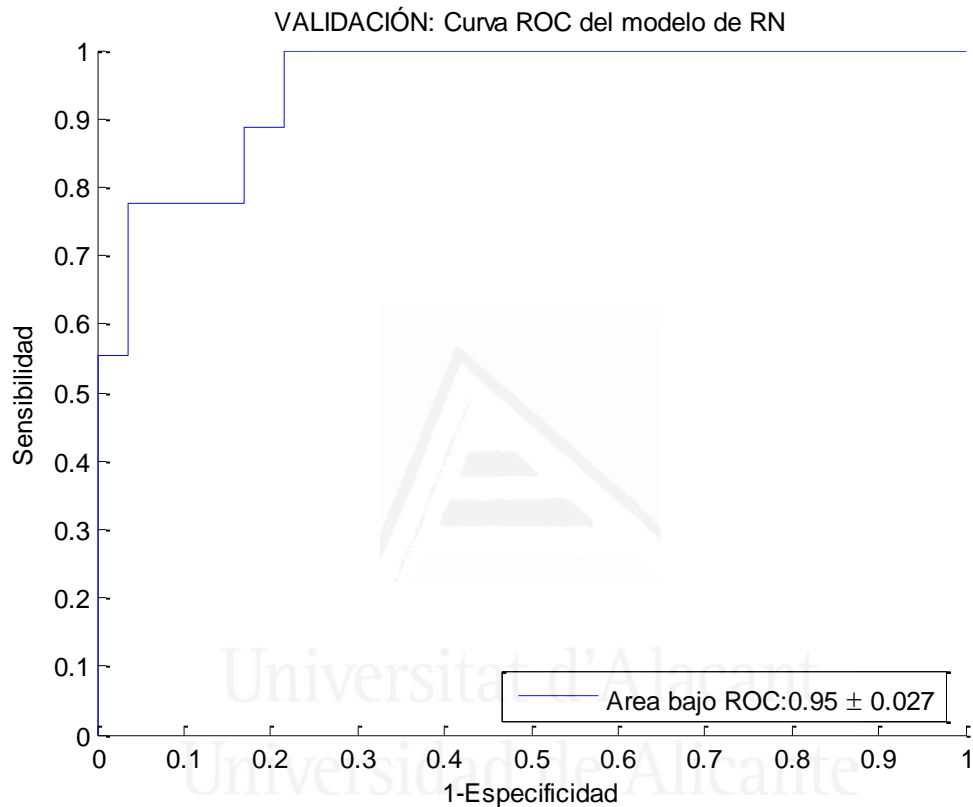
Muerte en UCIP estimada con el índice "PRISM":
Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana



2.2.1.- Evaluación clásica de la exactitud diagnóstica:

2.2.1.1.- Capacidad discriminante:

a) Estudio de la Curva ROC:



Área bajo la curva

Área	Error típico	Signif. Asintótica ^b	Intervalo de confianza asintótico al 95%		Método
			Límite inferior	Límite superior	
0,9552	0,0230	0,000	0,9102	1,0002	DeLong
	0,0485	0,000	0,8602	1,0501	Hanley & McNeil

b Hipótesis nula: área verdadera = 0,5

b) Índices de exactitud diagnóstica:

Pronosticado RL	Observado		
	Muerto	Vivo	
Positivo (Muerto)	8	22	30
Negativo (Vivo)	1	149	150
	9	171	180

		IC 95%
Sensibilidad	88,9%	68,4% a 109,4%
Especificidad	87,1%	82,1% a 92,2%
Valor predictivo positivo	26,7%	10,8% a 42,5%
Valor predictivo negativo	99,3%	98,0% a 100,6%
Proporción de falsos positivos	12,9%	7,8% a 17,9%
Proporción de falsos negativos	11,1%	-9,4% a 31,6%
Acierto	87,2%	82,3% a 92,1%
Odds ratio diagnóstica	53,15	13,42 a 219,5
Índice J de Youden	0,8	
Razón de Verosimilitudes + [LR(+)]	6,91	4,39 a 10,87
Razón de Verosimilitudes - [LR(-)]	0,13	0,02 a 0,81
Probabilidad pre-prueba (Prevalencia)	5,0%	

2.2.1.2.- Calibración del modelo:

Tabla de contingencias para la prueba de Hosmer y Lemeshow

Estratos de Riesgo	Exitus (en ese ingreso)				Total de niños
	Vivo		Muerto		
	Observado	Esperado	Observado	Esperado	
1	49	48.8808	0	0.1192	49
2	5	4.9859	0	0.0141	5
3	29	28.8820	0	0.1180	29
4	8	7.9597	0	0.0403	8
5	19	18.8812	0	0.1188	19
6	20	19.8151	0	0.1849	20
7	20	19.7342	1	1.2658	21
8	10	9.7832	1	1.2168	11
9	11	11.7899	7	6.2101	18

Prueba de Hosmer y Lemeshow

Ji-Cuadrado	Grados Libertad	Significación
0.8585	7	0.9968

No se puede descartar la hipótesis nula de "bondad de ajuste"

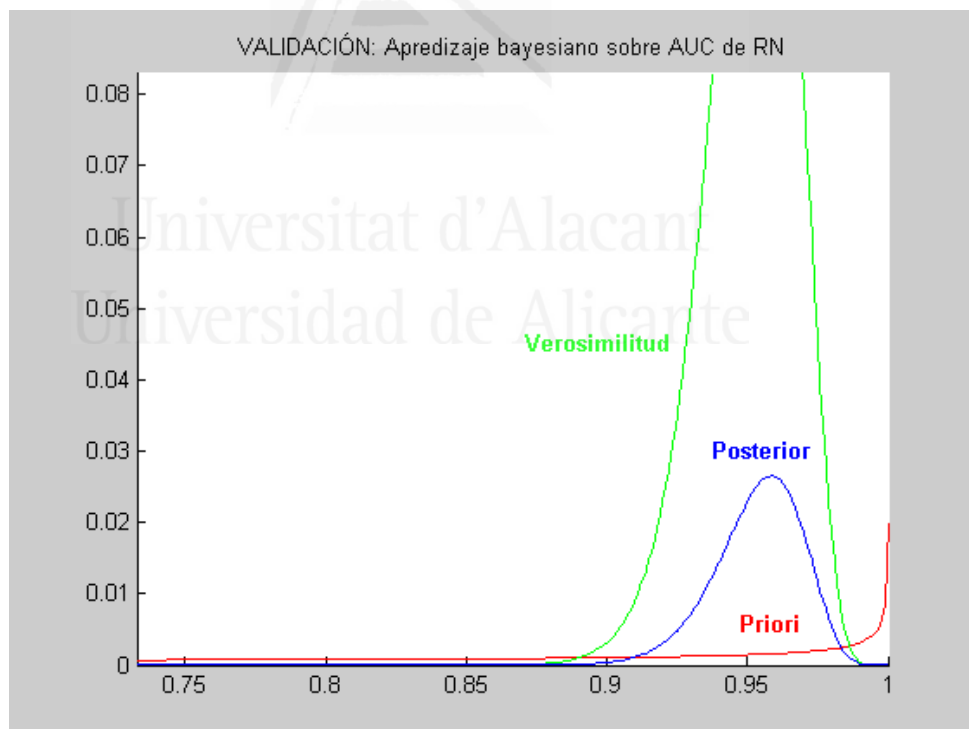
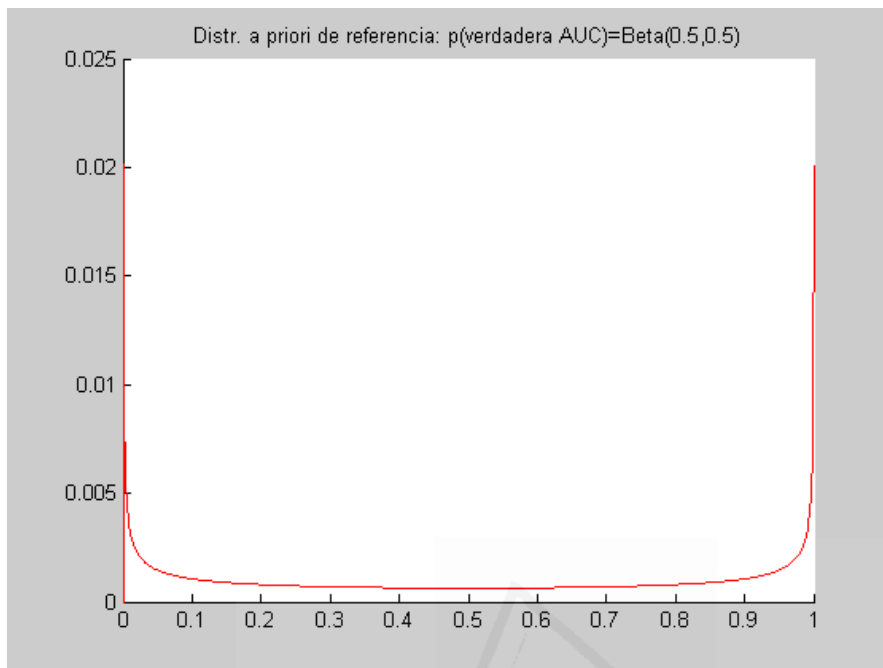
2.2.2.- Evaluación bayesiana de la exactitud diagnóstica:

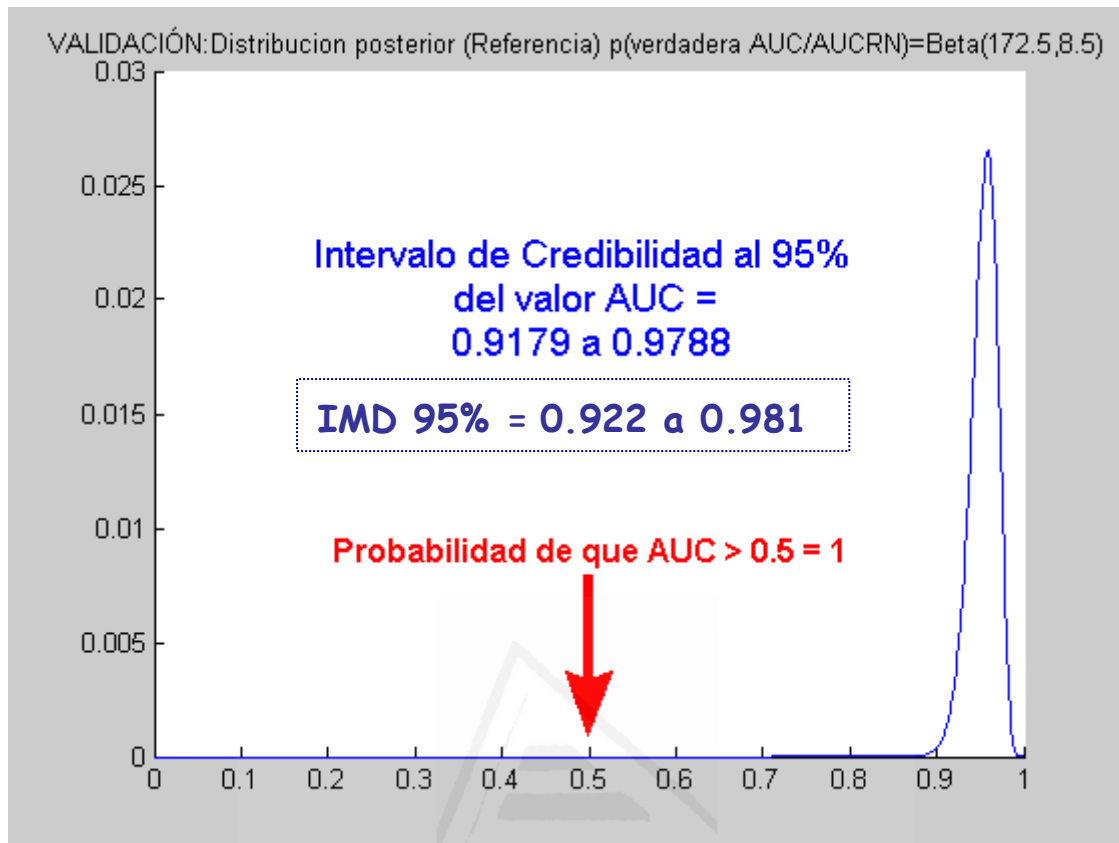
2.2.2.1.- Capacidad discriminante:

a) Estudio bayesiano de la Curva ROC:

Se considera que la AUC es una probabilidad, por lo que se modeliza con la distribución BETA. Para el análisis bayesiano de referencia, se toma como distribución a priori sobre el parámetro "verdadero valor de la AUC" la **Beta (1/2, 1/2)**. Por el estudio clásico y mediante el método trapezoidal, se sabe que el área bajo curva ROC es 0.9552, y que el tamaño muestral es $n = 180$, por lo que la función de verosimilitud conjugada de la Beta que modeliza el experimento es la Binomial ($r=172, n=180, AUC$).

Aplicando el teorema de Bayes, se obtiene que la distribución posterior sobre el parámetro "verdadero valor de la AUC" es la distribución **Beta (172.5, 8.5)**, cuya media es **AUC = 0.953**. Integrándola, podemos calcular que tenemos una probabilidad del 95% de que el "verdadero valor de la AUC" quede entre 0.9179 y 0.9788 (Intervalo de Credibilidad al 95%). Como es asimétrica, estimamos el IMD, que resulta ser **IMD 95% = 0.922 a 0.981**. La probabilidad de que el "verdadero valor de la AUC" sea mayor que 0.5 -esto es, que el modelo de regresión logística clasifique mejor que el azar- es, después de aprender de los datos del experimento, prácticamente 1.





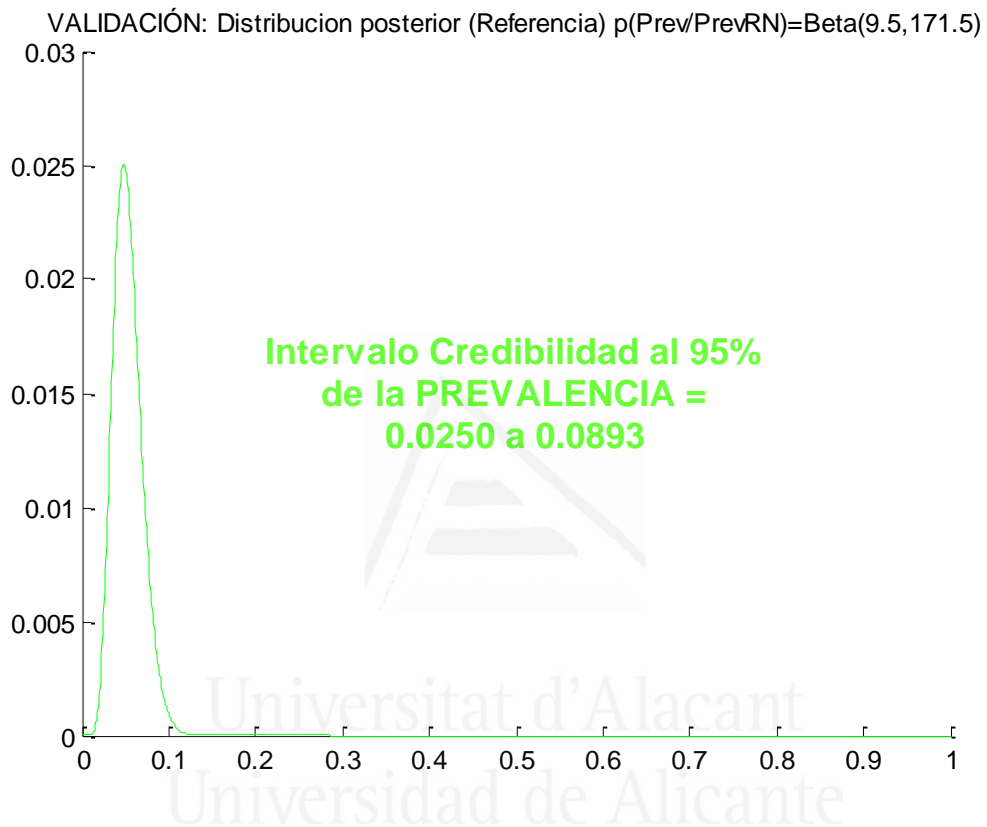
b) Índices bayesianos de exactitud diagnóstica:

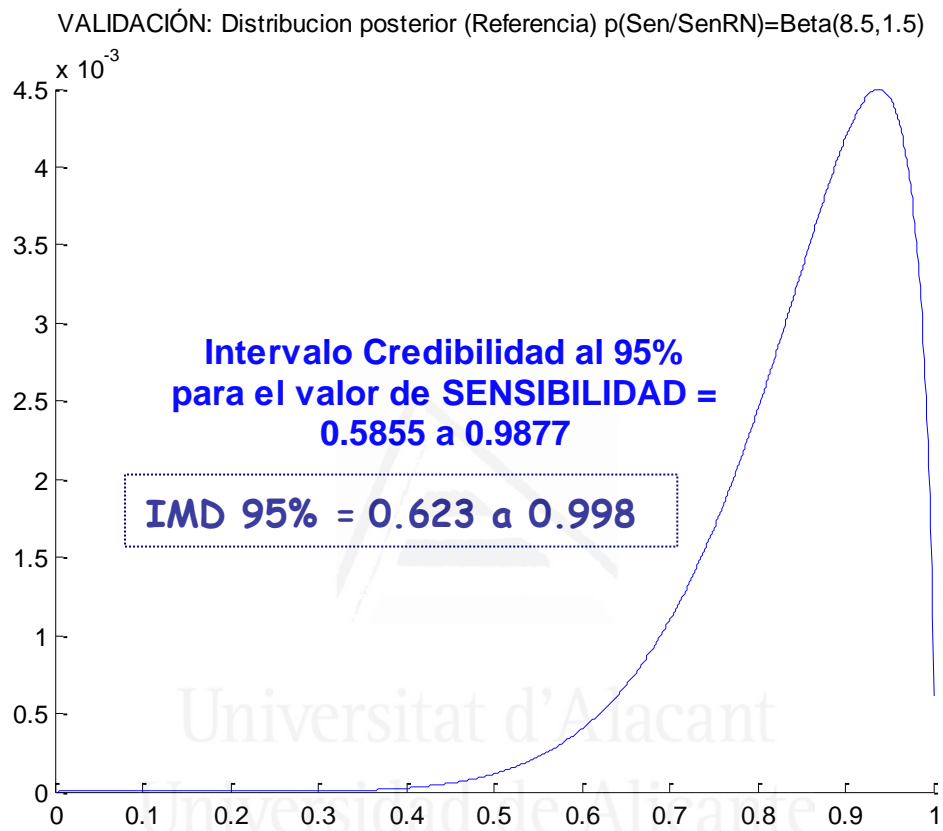
Tanto la Prevalencia como la Sensibilidad, la Especificidad y los Valores Predictivos son probabilidades, por lo que también se modelizan con la distribución BETA. Las Razones de Verosimilitud se modelizan con la distribución NORMAL, usando la transformación logarítmica. Para el análisis bayesiano de referencia, se toma como distribución a priori sobre ellas la **Beta (1/2, 1/2)**. En la tabla siguiente se detallan sus distribuciones de verosimilitud, sus distribuciones posteriores, y los Intervalos de Credibilidad y de Máxima Densidad al 95%. Los resultados se han obtenido aplicando el Teorema de Bayes.

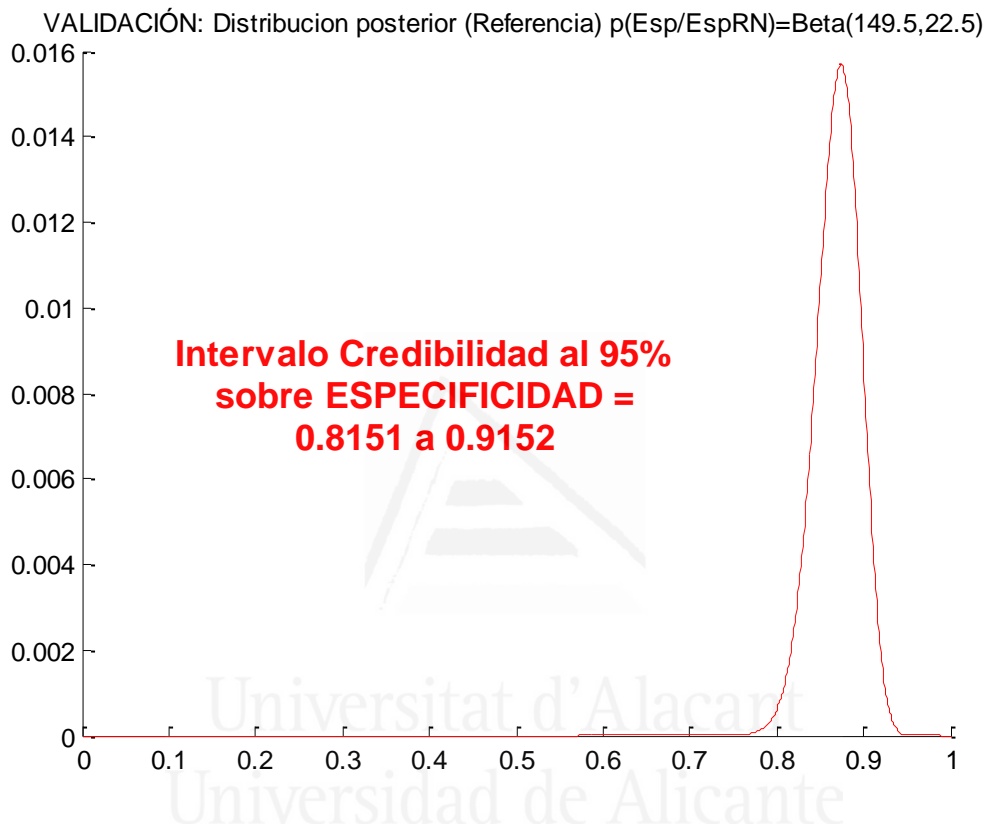
Pronosticado RL	Observado		
	Muerto	Vivo	
Positivo (Muerto)	8	22	30
Negativo (Vivo)	1	149	150
	9	171	180

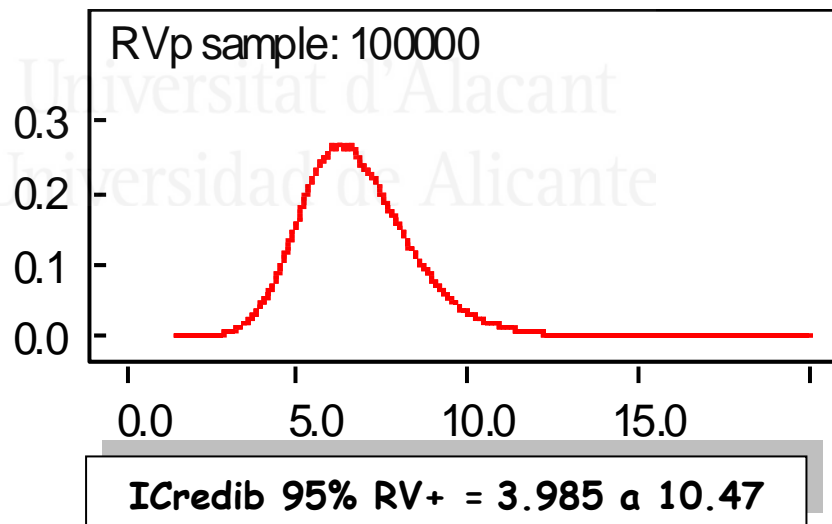
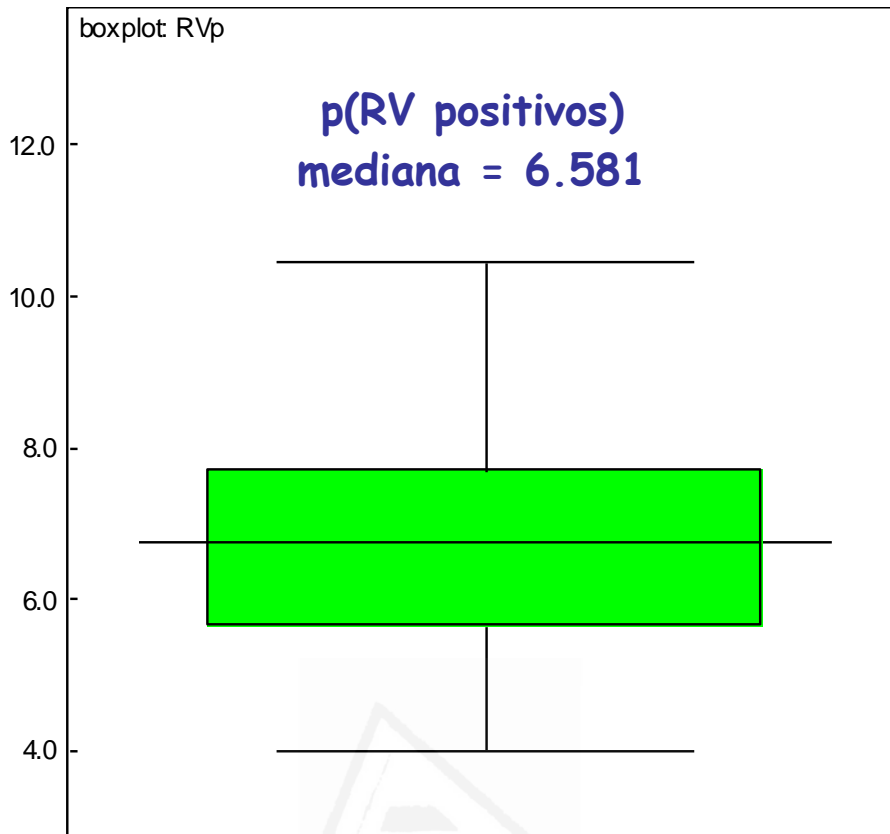
	Priori	Verosimilitud	Posteriori
Prevalencia	Be(1/2,1/2)	Bi(r=9,n=180,P)	Be(9.5, 171.5)
Sensibilidad		Bi(r=8,n=9,S)	Be(8.5, 1.5)
Especificidad		Bi(r=149,n=171,E)	Be(149.5, 22.5)
VP Positivos		Bi(r=8,n=30,VPP)	Be(8.5, 22.5)
VP Negativos		Bi(r=149,n=150,VPN)	Be(149.5, 1.5)
Raz Veros +	RVpos = Sens/(1 - Esp)		
Raz Veros -	RVneg = (1 - Sens)/Esp		

	Media (Post)	Int Credib 95%	IMD 95%
Prevalencia	5.00 %	2.5 a 8.93 %	2.2 a 8.5 %
Sensibilidad	85.0 %	58.55 a 98.77 %	62.3 a 99.8 %
Especificidad	87.0 %	81.51 a 91.52 %	81.8 a 91.7 %
VP Positivos	27.0 %	13.5 a 44.1 %	12.6 a 43 %
VP Negativos	99.33 %	96.9 a 99.999 %	96.5 a 99.8 %
Raz Veros +	6.752	3.985 a 10.47	
Raz Veros -	0.1731	0.01391 a 0.4796	



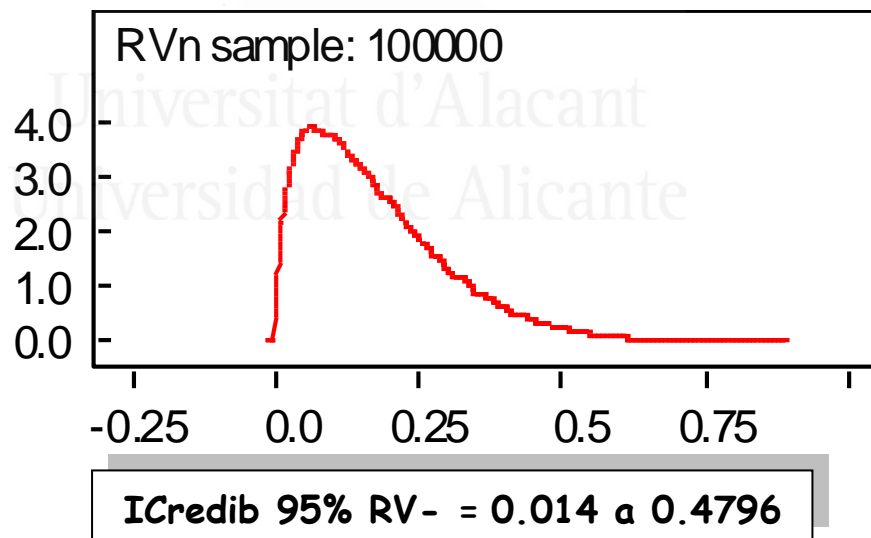
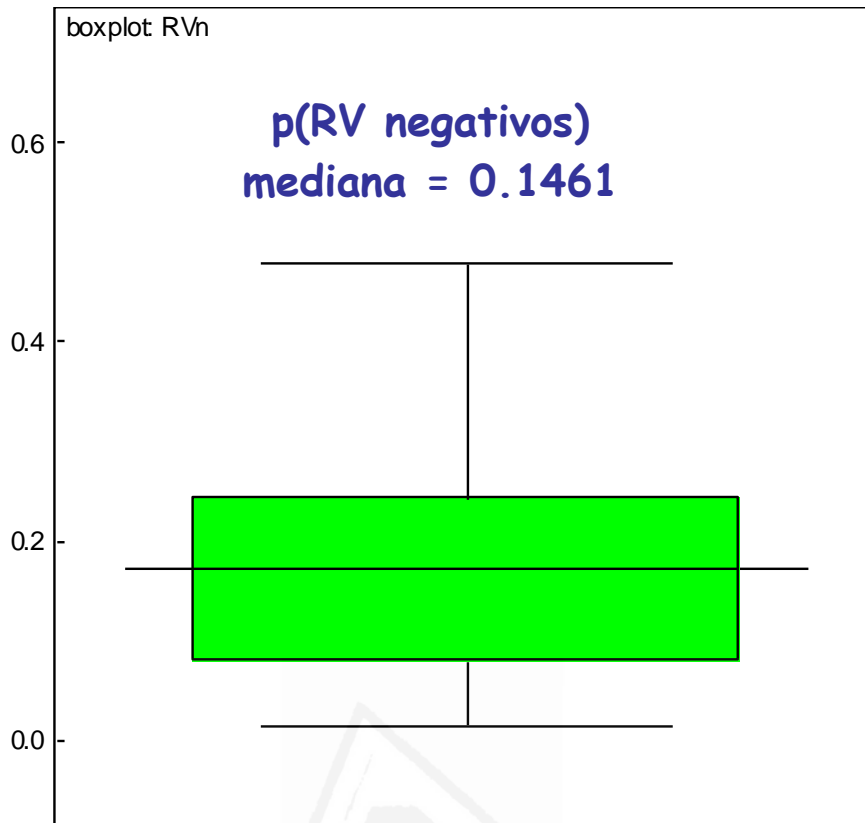






Si consideramos que una RV+ es fuerte cuando es superior a 5 (TABLA 3 y TABLA 4, pag.43), puede calcularse también la probabilidad de que la Red Neuronal presente una RV+ > 5 y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV+ > 5) = 0.8738$$



Si consideramos que una RV- es fuerte cuando es inferior a 1/5 (TABLA 3 y TABLA 4, pag.43), puede calcularse la probabilidad de que la Red Neuronal presente una $RV- < 1/5$ y sea por tanto un buen método diagnóstico. Así, integrando, tenemos:

$$P(RV- < 1/5) = 0.655$$

2.2.2.2.- Calibración del modelo

Tabla de contingencias para la prueba de Hosmer y Lemeshow

Estratos de Riesgo	Exitus (en ese ingreso)				Total de niños
	Vivo		Muerto		
	Observado	Esperado	Observado	Esperado	
1	49	48.8808	0	0.1192	49
2	5	4.9859	0	0.0141	5
3	29	28.8820	0	0.1180	29
4	8	7.9597	0	0.0403	8
5	19	18.8812	0	0.1188	19
6	20	19.8151	0	0.1849	20
7	19	18.7545	2	2.2455	21
8	11	10.7406	0	0.2594	11
9	11	12.1766	7	5.8234	18

a) Usando el estimador bayesiano convencional:

Entropía Observados = 2.7789 bits

Entropía Esperados = 2.9866 bits

Entropía Relativa Obs a Esp: $D(o||e) = 0.0134$ bits

Entropía Relativa Esp a Obs: $D(e||o) = 0.0138$ bits

Divergencia de Jeffreys $J(o,e) = 0.0271$ bits

Discrepancia Intrínseca $\delta(o,e) = 0.0134$ bits

El modelo de RN queda a una distancia despreciable de la observada

b) Usando el estimador bayesiano intrínseco:

Entropía Observados = 2.7608 bits

Entropía Esperados = 2.9842 bits

Entropía Relativa Obs a Esp: $D(o||e) = 0.0153$ bits

Entropía Relativa Esp a Obs: $D(e||o) = 0.0161$ bits

Divergencia de Jeffreys $J(o,e) = 0.0314$ bits

Discrepancia Intrínseca $\delta(o,e) = 0.0153$ bits

El modelo de RN queda a una distancia despreciable de la observada



Universitat d'Alacant
Universidad de Alicante

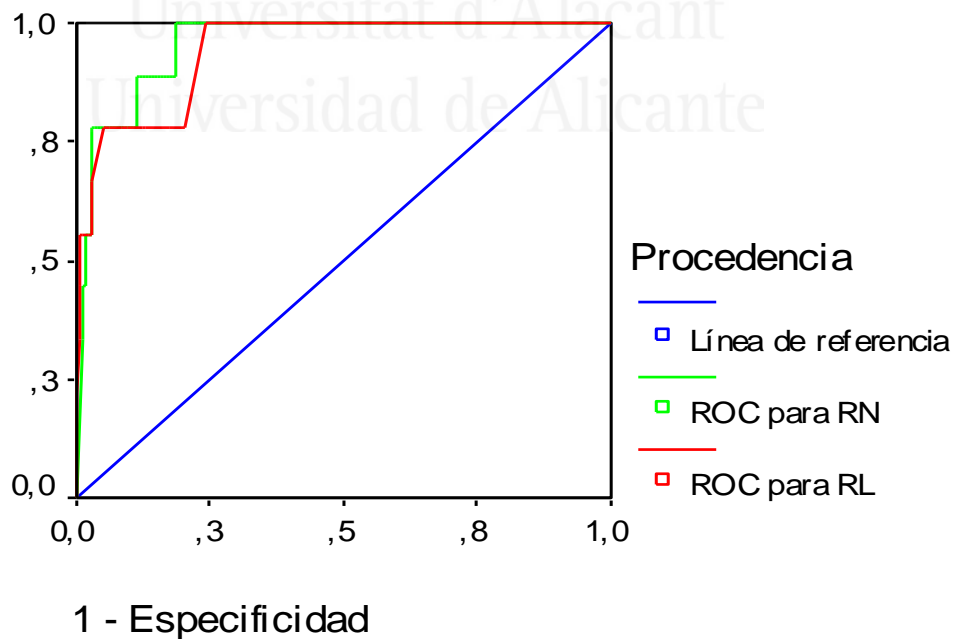
3.- 3ª FASE: COMPARACIÓN DE LA EXACTITUD DIAGNÓSTICA ENTRE AMBOS TEST

3.1.- EVALUACIÓN CLÁSICA DE LA DIFERENCIA EN LA EXACTITUD DIAGNÓSTICA:

3.1.1.- Diferencia en la Capacidad Discriminante:

3.1.1.1.- Análisis del Área bajo las curvas ROC (AUC):

Curva ROC



Los segmentos diagonales son producidos por los empates.

Quando se trata de curvas ROC correlacionadas, la comparación clásica se hace con la prueba no paramétrica de De Long¹⁴¹.

Área bajo las curvas ROC

Curva	Área (AUC)	Error típico	Intervalo de confianza asintótico al 95%		Método
			Límite inferior	Límite superior	
Red Neuronal	0,9552	0,0230	0,9102	1,0002	DeLong
Regr. Logística	0,9412	0,0325	0,8774	1,0049	

Prueba de DeLong de la homogeneidad de las áreas bajo las curvas ROC (áreas correlacionadas)

Ji-Cuadrado	Grados de Libertad	Valor p
0,8862	1	0,3465

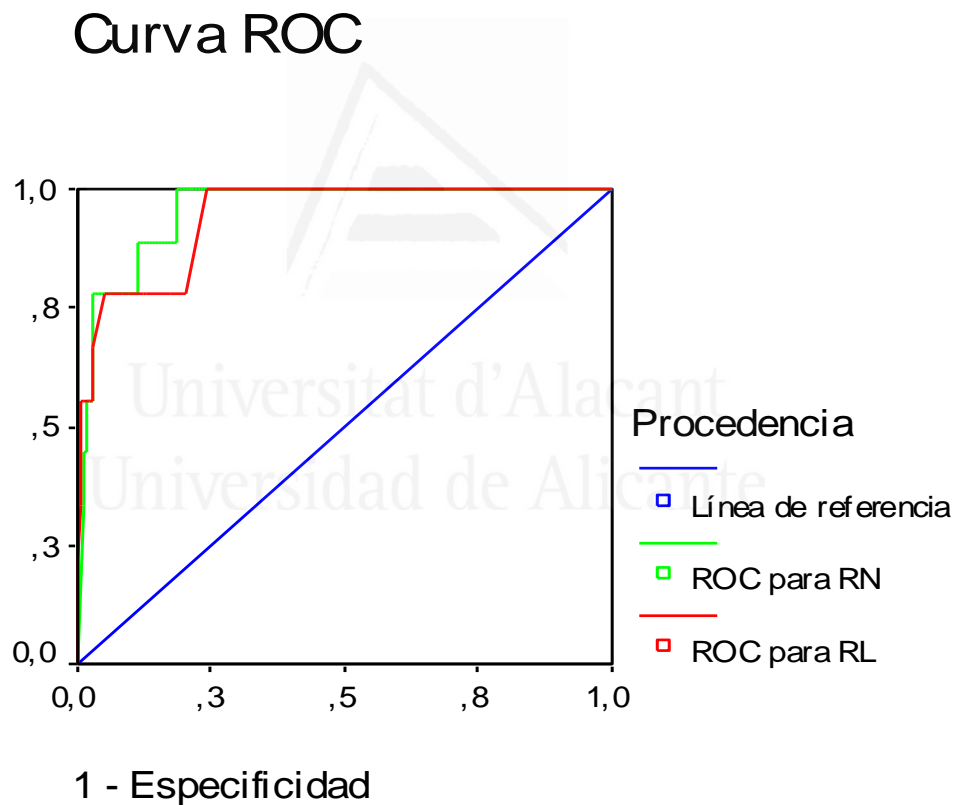
3.1.2.- Diferencia en la Calibración:

Clásicamente, no se utiliza ningún test capaz de encontrar diferencias estadísticamente significativas en la calibración de dos métodos.

3.2.- PROPUESTA DE EVALUACIÓN BAYESIANA DE LA DIFERENCIA EN LA EXACTITUD DIAGNÓSTICA:

3.2.1.- Diferencia en la Capacidad Discriminante:

3.2.1.1.- Análisis bayesiano del Área bajo las curvas ROC (AUC):



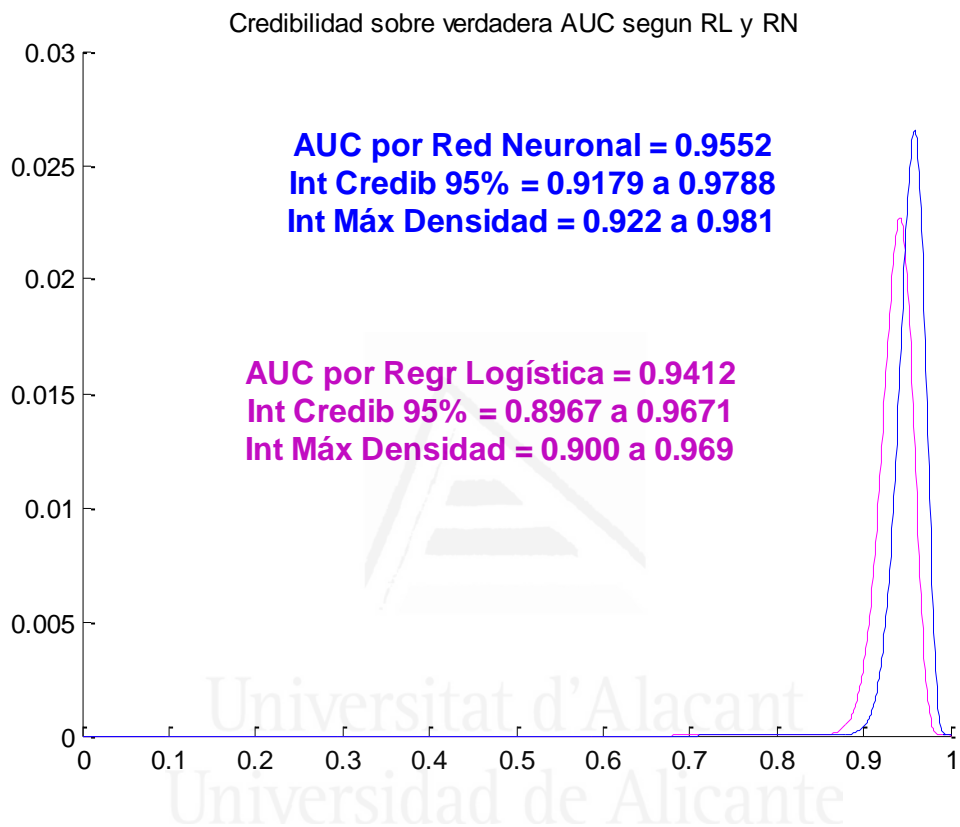
Los segmentos diagonales son producidos por los empates.

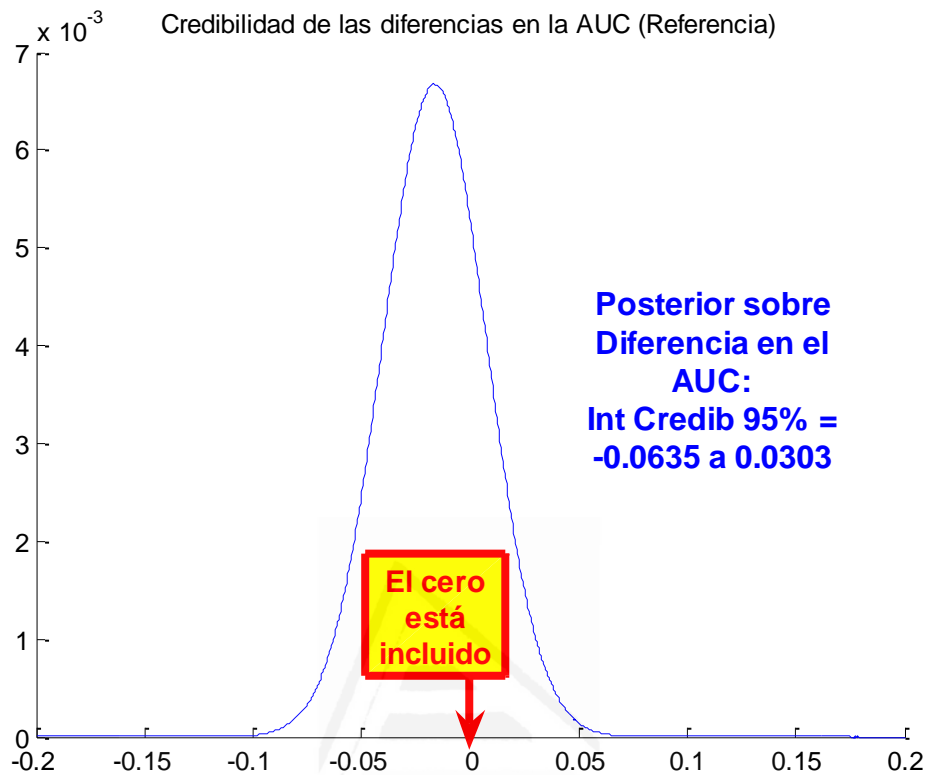
a) Utilizando la aproximación Normal:

Ya que el tamaño muestral es grande (n = 180), podemos asumir la aproximación Normal a la distribución Beta. En la tabla siguiente se detallan las distribuciones Beta posteriores de cada AUC, sus aproximaciones Normales, y la distribución posterior Normal sobre la diferencia en AUC, así como sus intervalos de Credibilidad al 95%.

		Distribución posterior sobre AUC	
		Regresión Logística	Red Neuronal
Distribución Beta		Beta (169.5, 11.5)	Beta (172.5, 8.5)
Media		0.9365	0.9530
Varianza		3.2692e-4	2.4591e-4
Int Credib al 95%	Lím Inferior	0.8967	0.9179
	Lím Superior	0.9671	0.9788
Aproximación Normal		N (0.9365, $\sqrt{3.2692e-4}$)	N (0.9530, $\sqrt{2.4591e-4}$)
Distribución posterior sobre Diferencia entre AUCs (AUC/RL-AUC/RN)		$N(0.9365, \sqrt{3.2692e-4}) - N(0.9530, \sqrt{2.4591e-4}) =$ $N([0.9365-0.9530], \sqrt{[3.2692e-4 + 2.4591e-4]}) =$ N(-0.0166, 0.0239)	
Media		-0.0166	
Varianza		5.7283e-4	
Desviación Típica		0.0239	
Int Credib al 95%	Lím Inferior	-0.0635	
	Lím Superior	0.0303	

El cero (idéntica capacidad de discriminación) es un valor posible





Integrando esta función podemos aproximar la probabilidad de que la RL tenga mayor capacidad discriminante que la RN, y viceversa. Así:

- Probabilidad de que RL sea más discriminante que la RN:

$$P(\text{AUC/RL} > \text{AUC/RN}) = P(\text{Dif en AUCs} > 0) = 0.2437$$

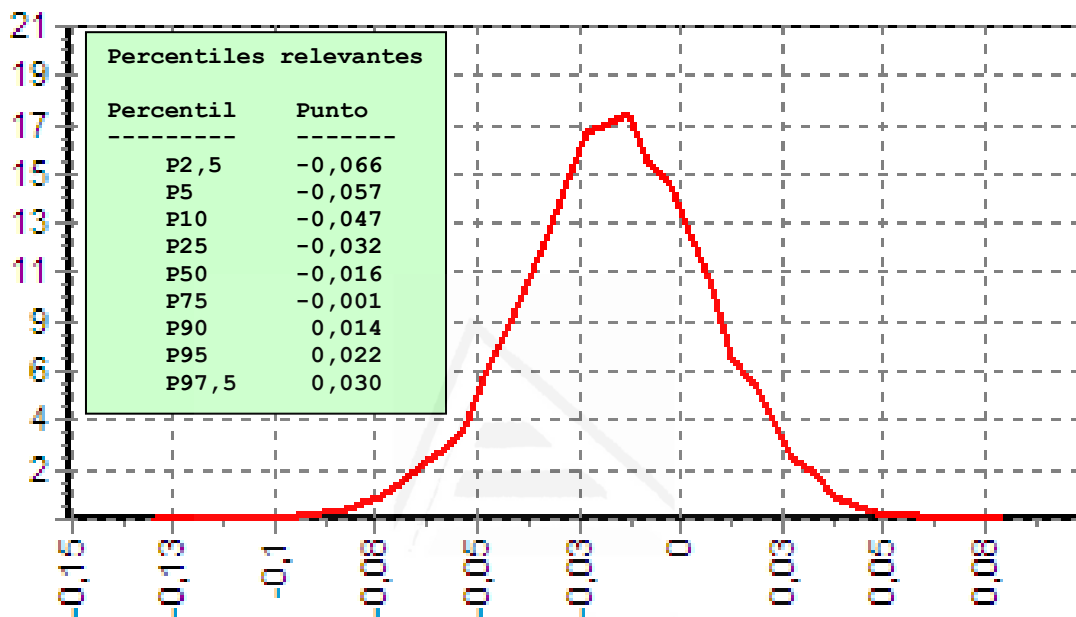
- Probabilidad de que la RN sea más discriminante que la RL:

$$P(\text{AUC/RL} < \text{AUC/RN}) = P(\text{Dif en AUCs} < 0) = 0.7563$$

Con ello podemos concluir que, aproximadamente, **es (3 veces) más probable que la RN tenga mayor capacidad de discriminación.**

b) Utilizando simulación con técnica de Monte Carlo:

Mediante técnica de simulación de Monte Carlo, utilizando 10000 simulaciones, podemos representar la distribución empírica a posteriori de referencia de las diferencias en AUC entre la RL y la RN. En la siguiente figura se representa dicha distribución con sus percentiles relevantes:



De esta manera también podemos, mediante integración, aproximar la probabilidad de que la RL tenga mayor capacidad discriminante que la RN, y viceversa. Así:

- Probabilidad de que RL sea más discriminante que la RN:

$$P(\text{AUC/RL} > \text{AUC/RN}) = P(\text{Dif en AUCs} > 0) = 0.242$$

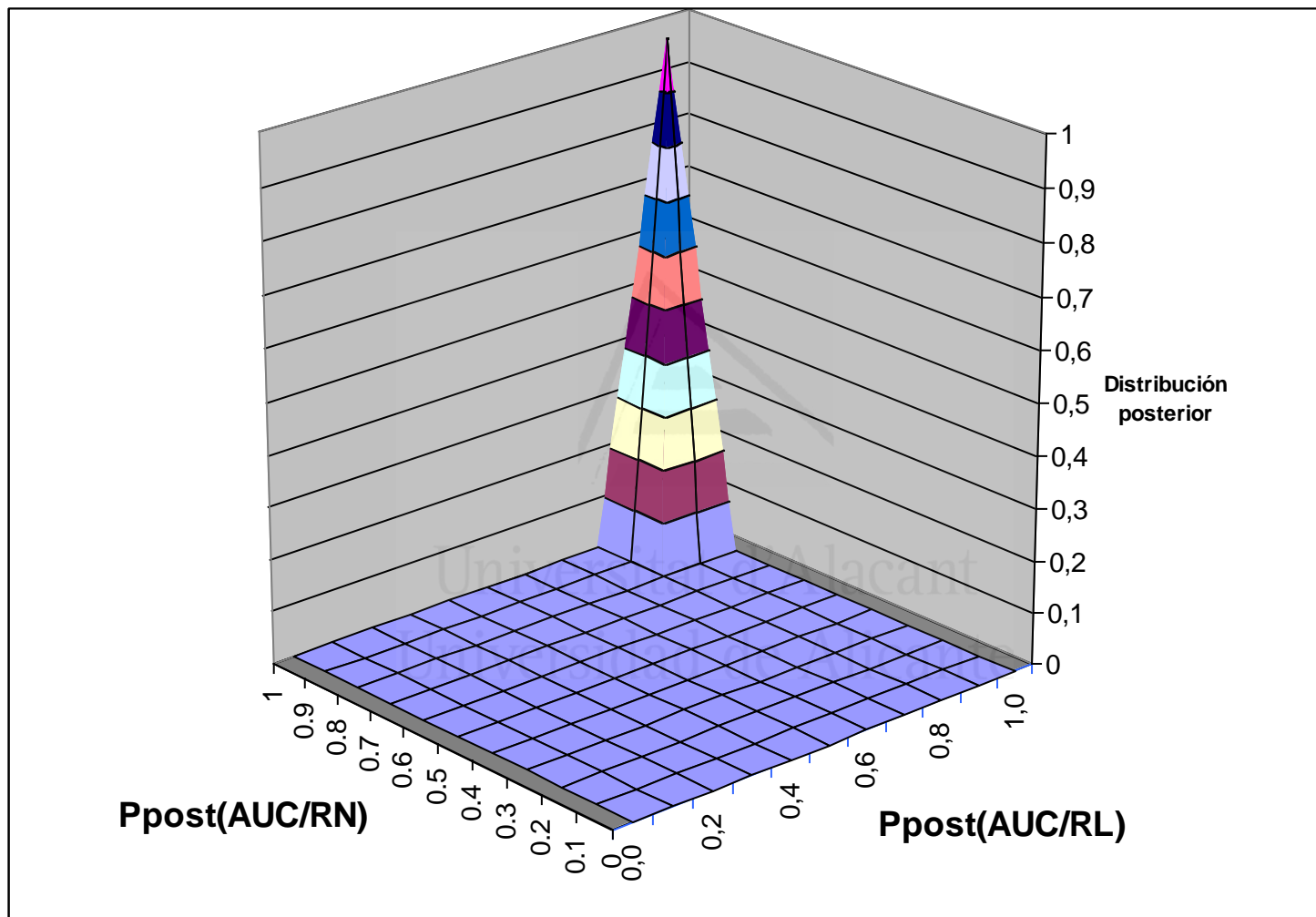
- Probabilidad de que la RN sea más discriminante que la RL:

$$P(\text{AUC/RL} < \text{AUC/RN}) = P(\text{Dif en AUCs} < 0) = 0.758$$

De nuevo podemos concluir que, aproximadamente, **es (3 veces) más probable que la RN tenga mayor capacidad de discriminación.**

Muerte en UCIP estimada con el índice "PRISM":
Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana

c) Utilizando el método gráfico discreto de la parrilla de Berry^{155,156}:



La diagonal queda incluida: ambos métodos tienen la misma capacidad discriminante

Muerte en UCIP estimada con el índice "PRISM":
 Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
 y una red neuronal artificial. Una propuesta bayesiana_____

Sumando las casillas, tenemos:

difAUC(RN-RL)	≤ -0.1	-0.067	-0.033	0	0.033	0.067	0.10	≥ 0.13
Probab. posterior	0	0.0051	0.0825	0.3928	0.4102	0.1065	0.0028	0.0001
Probab. Acumulada	0	0.0052	0.0876	0.4804	0.8906	0.9971	0.9999	1

Con esto podemos concluir:

- Probabilidad de que RL sea más discriminante que la RN:

$$P(\text{difAUC}(\text{RN-RL}) < 0) = 1 - 0.8906 = 0.1094$$

- Probabilidad de que la RN sea más discriminante que la RL:

$$P(\text{difAUC}(\text{RN-RL}) > 0) = 0.8906$$

- Probabilidad de que la capacidad discriminante de ambos métodos sea "aproximadamente" igual:

$$P(-0.033 < \text{difAUC}(\text{RN-RL}) < 0.033) = 0.4804$$

3.2.2.- Diferencia en la Calibración:

Nuestra propuesta para comparar la calibración de ambos métodos diagnósticos consiste en utilizar las medidas de entropía para caracterizar la proximidad entre cada uno de ellos y la distribución de lo observado. Así:

Distancia entrópica a lo observado

Predicción realizada por	Est. Bayesiano Convencional		Est. Bayesiano Intrínseco	
	Divergencia de Jeffreys $J(o, e)$	Discrepancia Intrínseca $\delta(o, e)$	Divergencia de Jeffreys $J(o, e)$	Discrepancia Intrínseca $\delta(o, e)$
Regr. Logística	0.048 bits	0.0235 bits	0.0582 bits	0.0280 bits
Red Neuronal	0.0271 bits	0.0134 bits	0.0314 bits	0.0153 bits

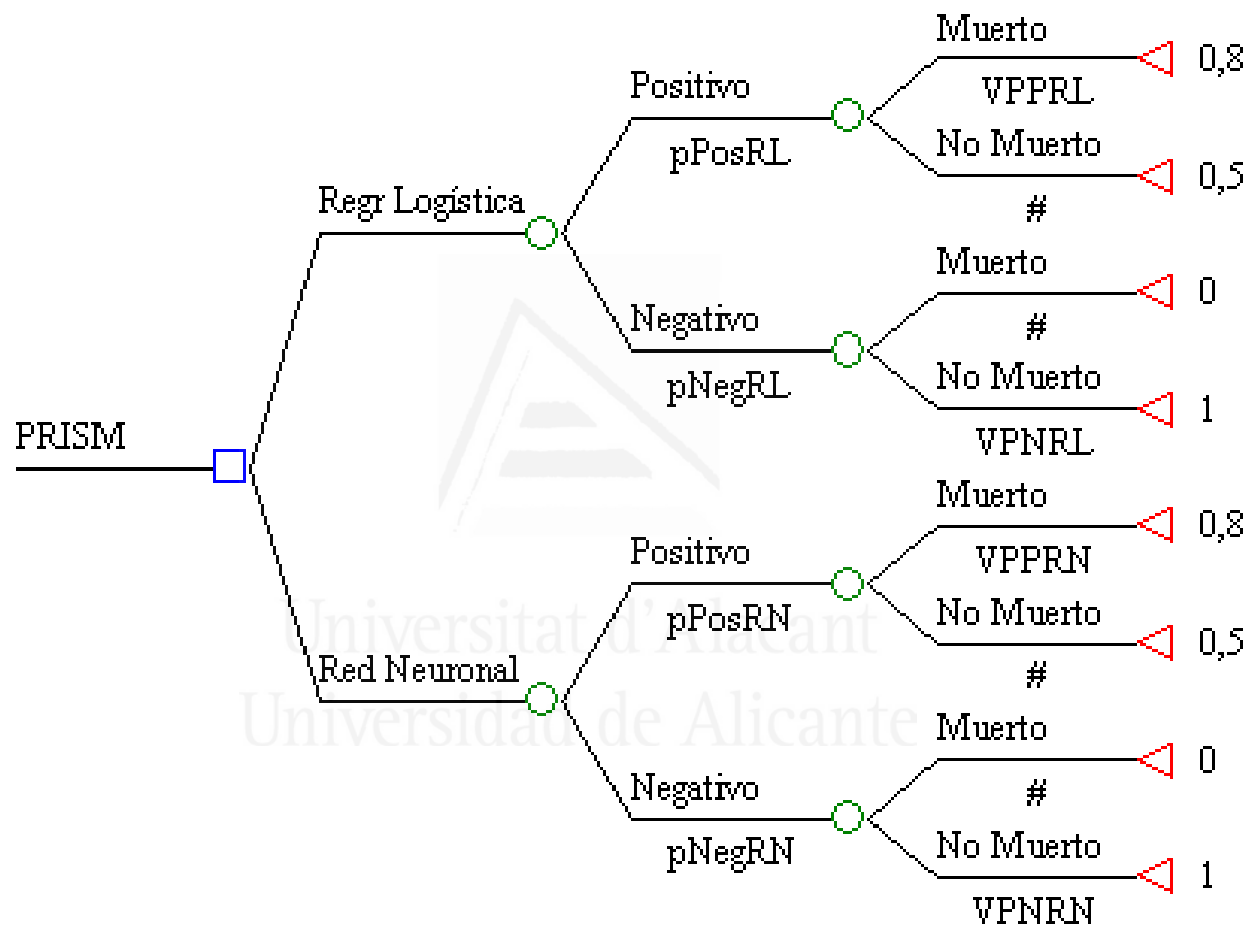
A pesar de que ambos métodos de diagnóstico se sitúan a una distancia despreciable de lo realmente observado, puede apreciarse que la calibración de la predicción realizada con la RN es muy superior: **la RL queda a una distancia prácticamente doble de la RN.**

3.3.- COMPARACIÓN MEDIANTE LA INCORPORACIÓN AL PROCESO DIAGNÓSTICO: ANÁLISIS DE DECISIÓN:

Desde el punto de vista clínico, la manera más útil de comparar ambos métodos diagnósticos es incorporarlos a un análisis formal de decisiones. Para ello, desde la perspectiva del médico que utiliza el test para informar a los padres del pronóstico de sus hijos, hemos asumido las siguientes utilidades:

- Verdaderos negativos: Máxima utilidad. Utilidad = 1
- Falsos negativos: Mínima utilidad. Utilidad = 0
- Verdaderos positivos: Utilidad = 0,8
- Falsos positivos: Utilidad = 0,5

Muerte en UCIP estimada con el índice "PRISM":
 Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
 y una red neuronal artificial. Una propuesta bayesiana



3.3.1.- Evaluación Clásica:

Decisión que maximiza utilidad esperada: **Red Neuronal : 0,924 Util Esp**

Regr Logística = 0,876 Util Esp

Positivo\0,278 :: 0,554 Util Esp

Muerto\0,180 :: 0,800 Utilidad

No Muerto\0,820 :: 0,500 Utilidad

Negativo\0,722 :: 1,000 Util Esp

Muerto\0,000 :: 0,000 Utilidad

No Muerto\1,000 :: 1,000 Utilidad

Red Neuronal = 0,924 Util Esp

Positivo\0,167 :: 0,580 Util Esp

Muerto\0,266 :: 0,800 Util Esp; P = 0,044

No Muerto\0,734 :: 0,500 Util Esp; P = 0,123

Negativo\0,833 :: 0,993 Util Esp

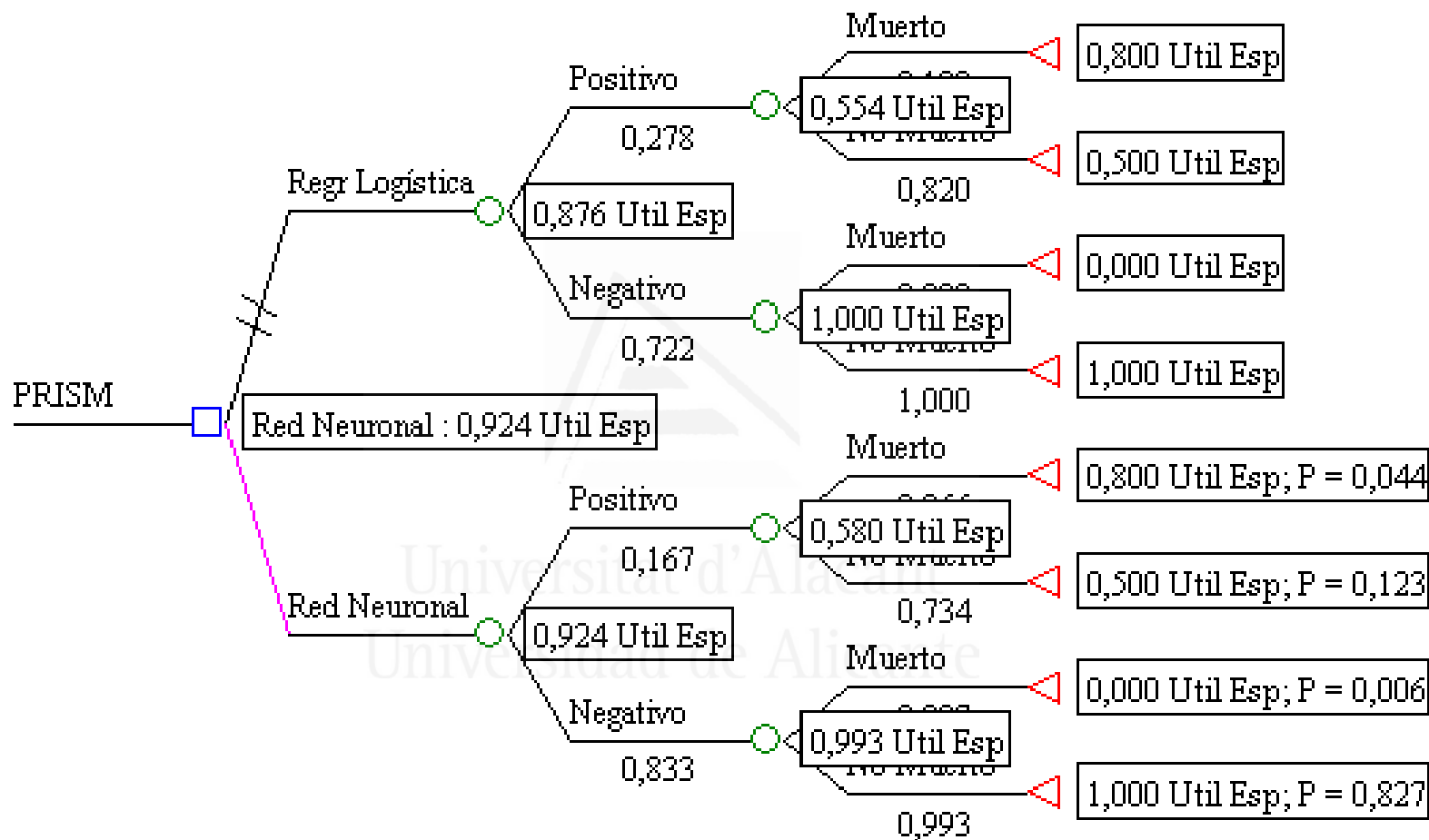
Muerto\0,007 :: 0,000 Util Esp; P = 0,006

No Muerto\0,993 :: 1,000 Util Esp; P = 0,827

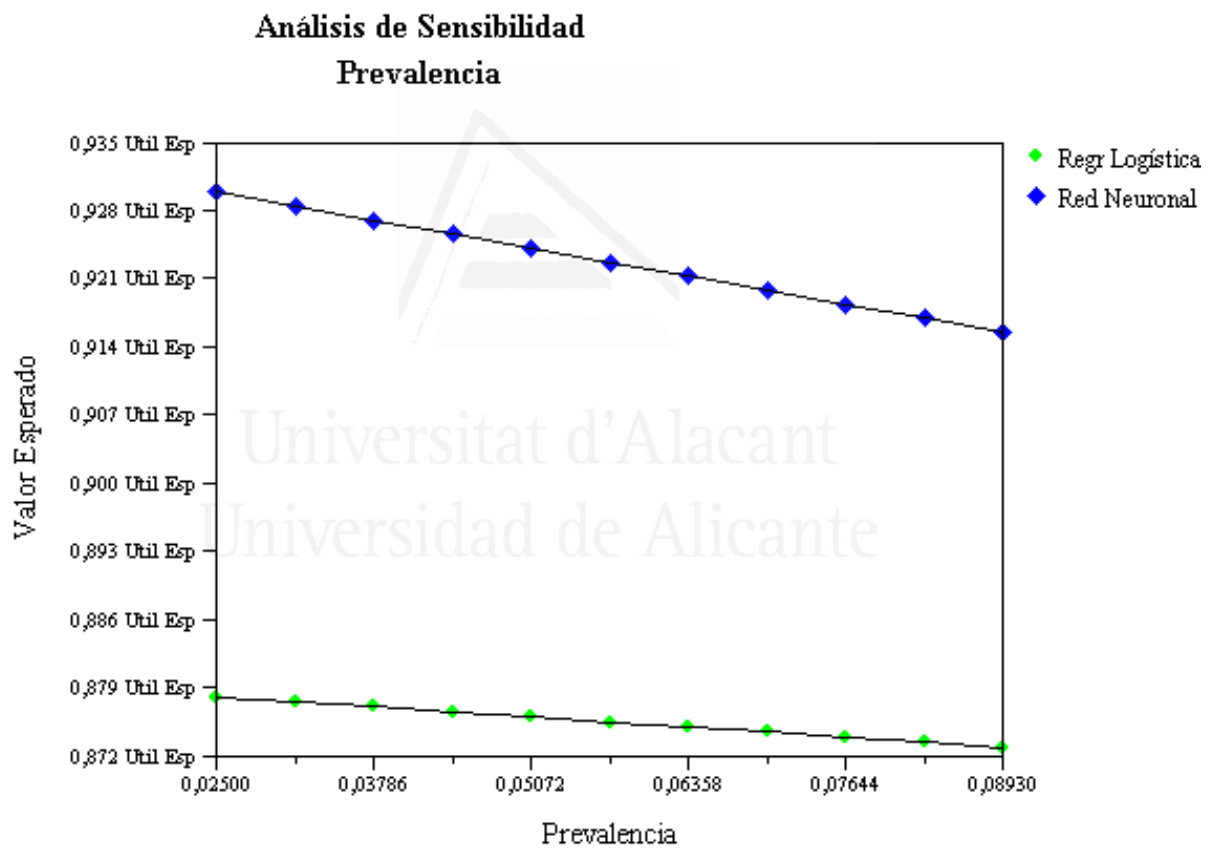
Universidad de Alicante

Muerte en UCIP estimada con el índice "PRISM":

Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística, y una red neuronal artificial. Una propuesta bayesiana

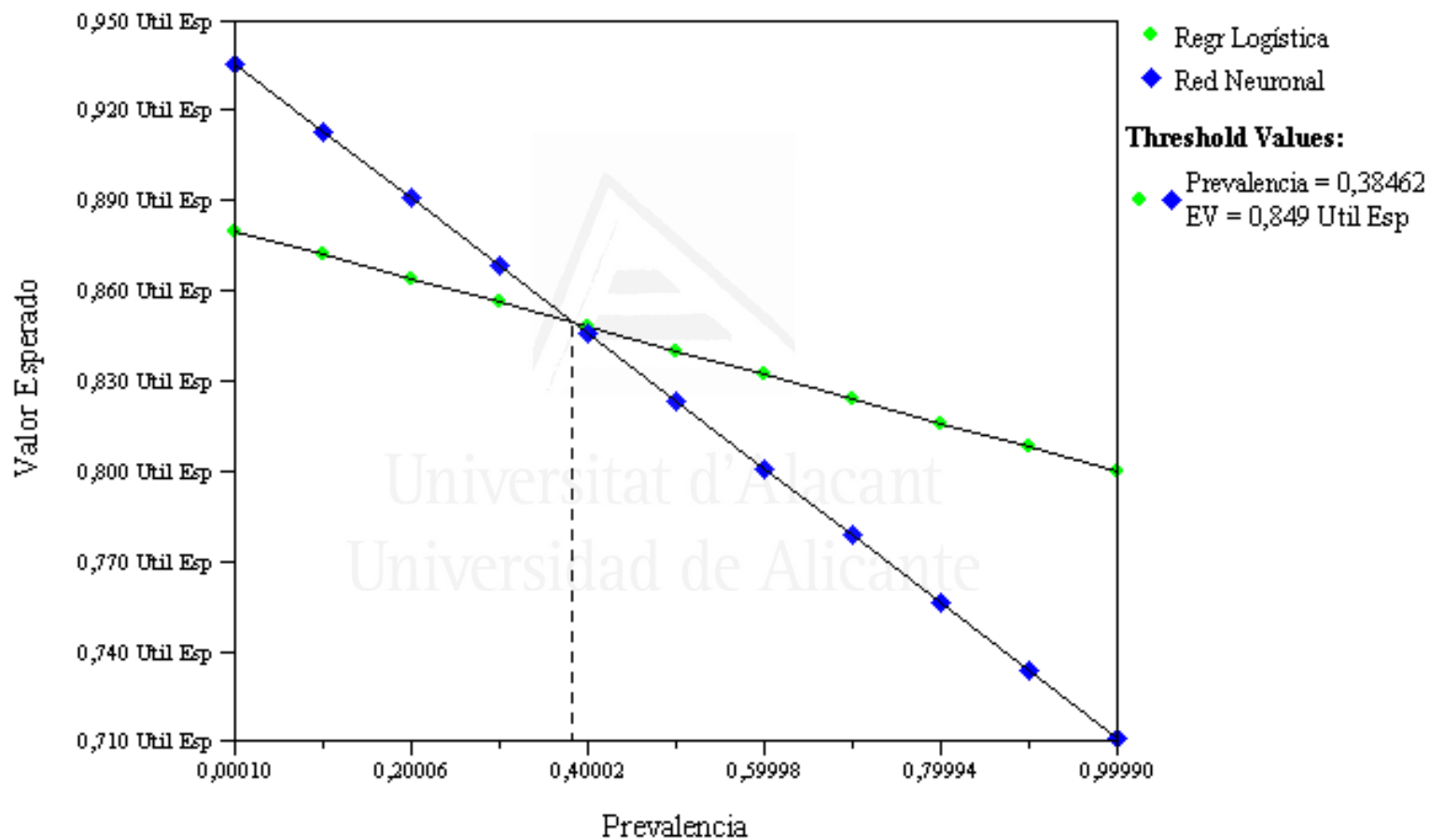


Tal como se observa en el siguiente análisis de sensibilidad, la decisión más adecuada es solicitar la RED NEURONAL durante todo el rango de prevalencias del Intervalo clásico de Confianza del 95% de la prevalencia de muerte en nuestra base de datos. Sólo cuando la prevalencia de muerte sea del 38,462% la decisión preferible es realizar la Regresión Logística.



Muerte en UCIP estimada con el índice "PRISM":
Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana

Análisis de Sensibilidad para Prevalencia



3.3.2.- Evaluación Bayesiana:

Decisión que maximiza utilidad esperada: **Red Neuronal : 0,922 Util Esp**

Regr Logística :: 0,874 Util Esp

Positivo\0,275 :: 0,552 Util Esp

Muerto\0,172 :: 0,800 Util Esp

No Muerto\0,828 :: 0,500 Util Esp

Negativo\0,725 :: 0,997 Util Esp

Muerto\0,003 :: 0,000 Util Esp

No Muerto\0,997 :: 1,000 Util Esp

Red Neuronal :: 0,922 Util Esp

Positivo\0,166 :: 0,577 Util Esp

Muerto\0,256 :: 0,800 Util Esp; P = 0,043

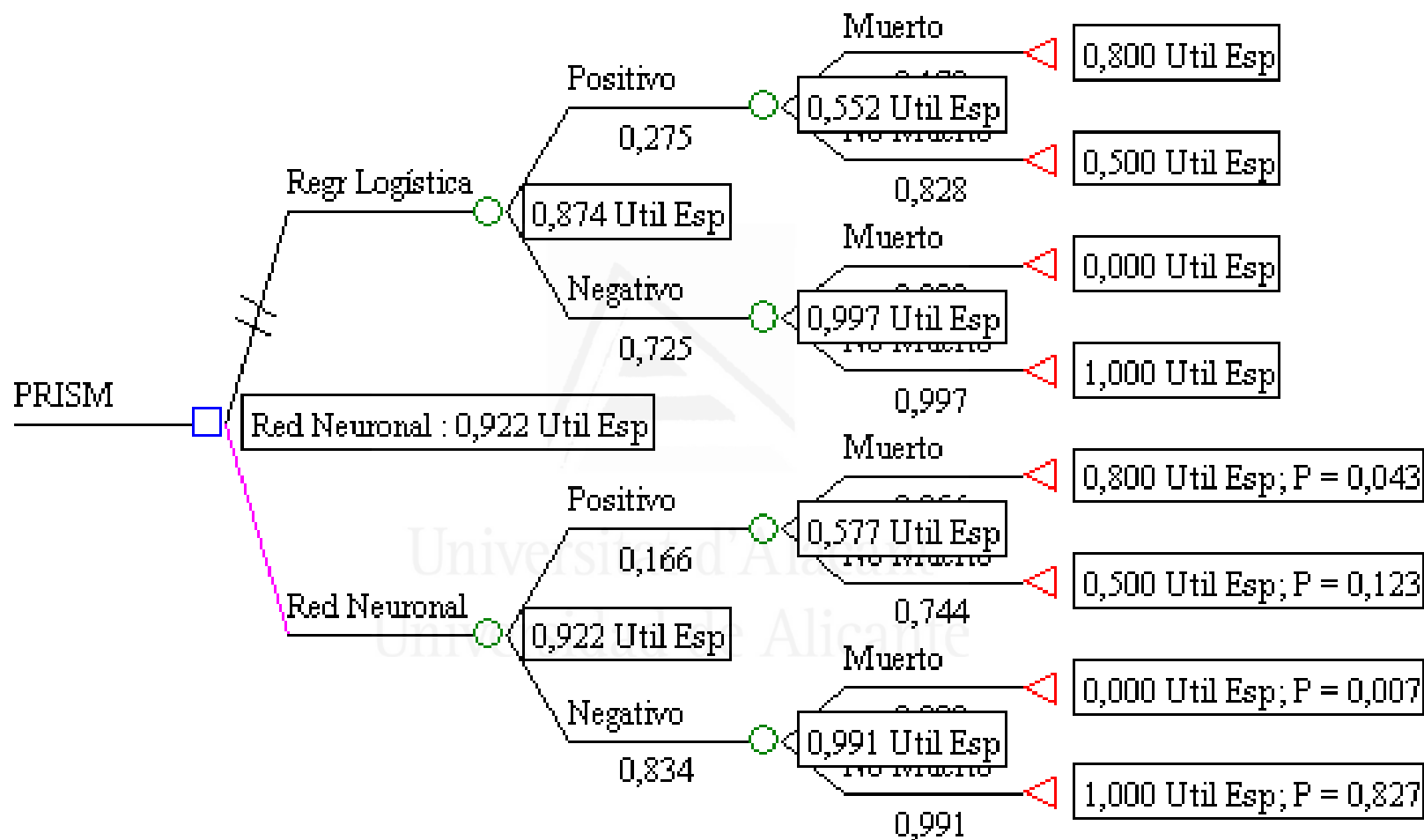
No Muerto\0,744 :: 0,500 Util Esp; P = 0,123

Negativo\0,834 :: 0,991 Util Esp

Muerto\0,009 :: 0,000 Util Esp; P = 0,007

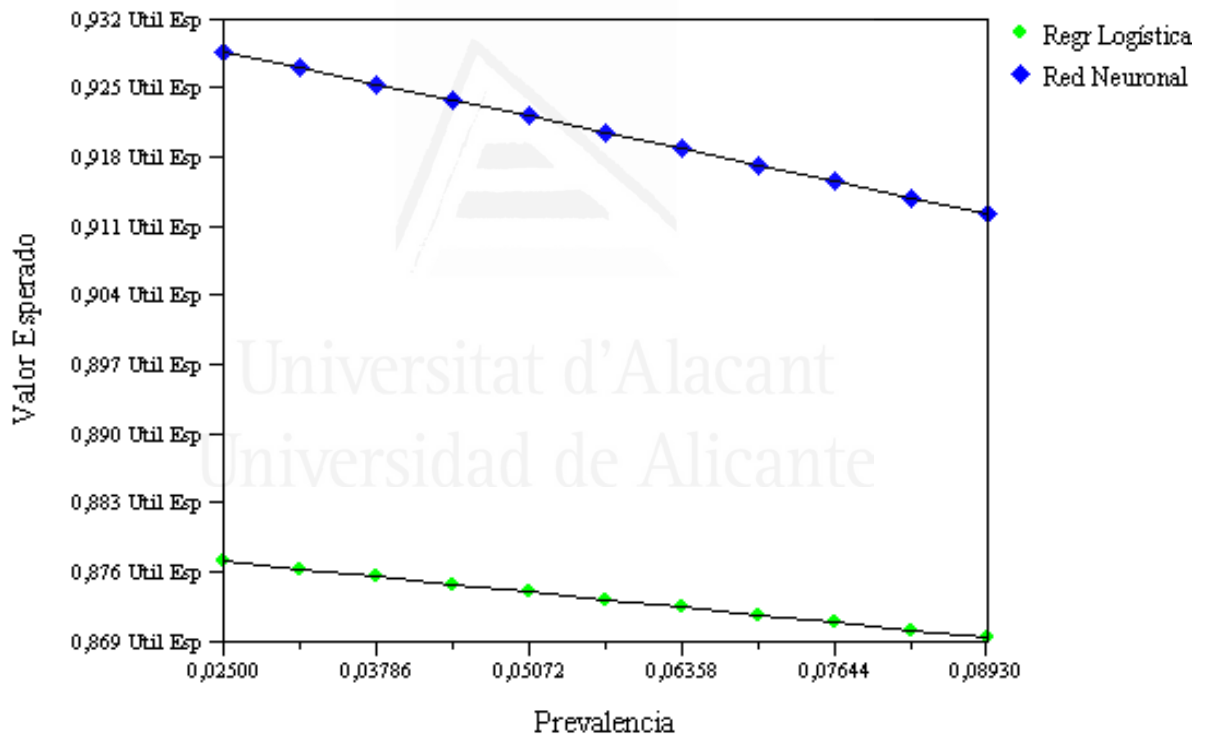
No Muerto\0,991 :: 1,000 Util Esp; P = 0,827

Muerte en UCIP estimada con el índice "PRISM":
 Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
 y una red neuronal artificial. Una propuesta bayesiana



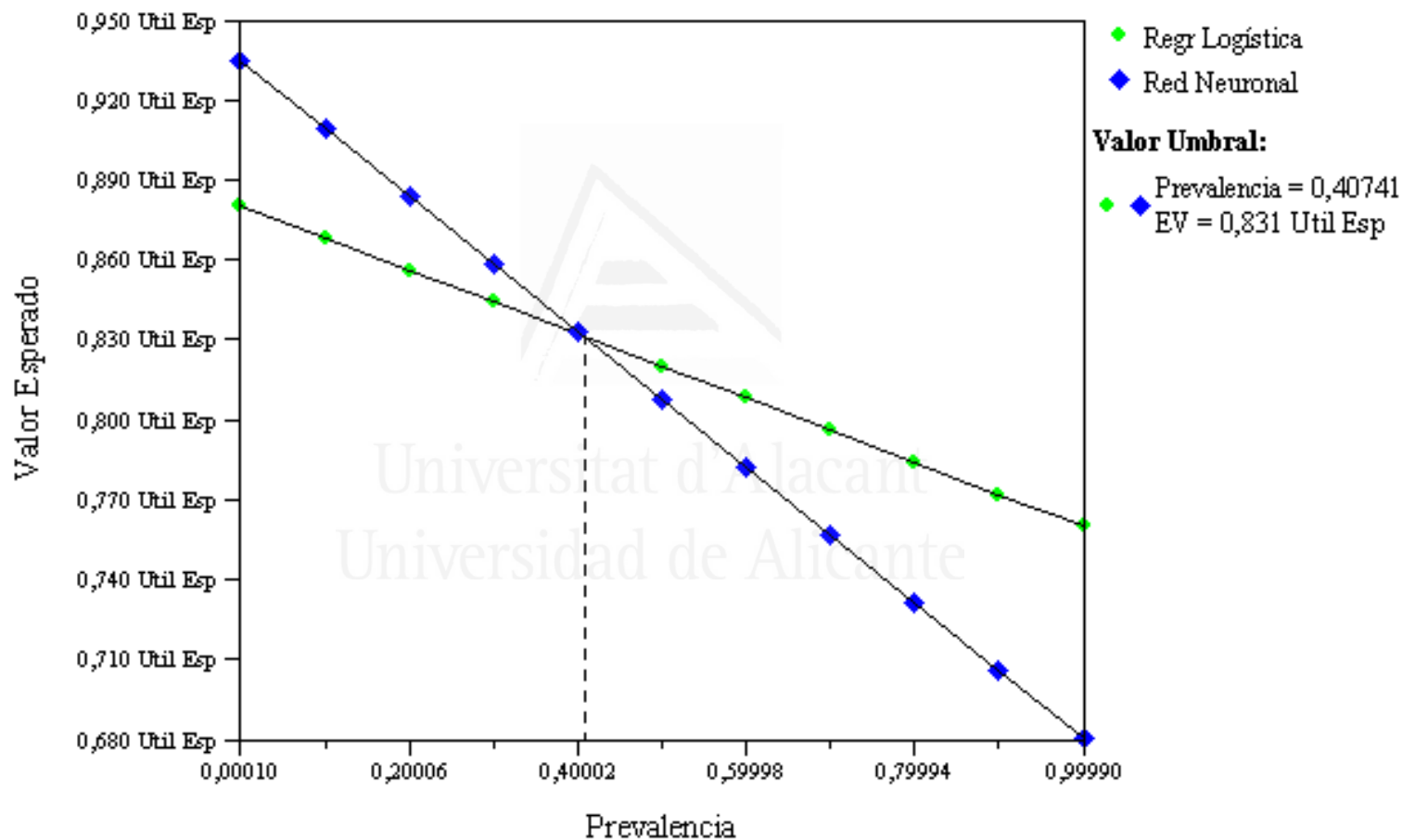
Desde el punto de vista bayesiano, el análisis de sensibilidad durante todo el rango de prevalencias del Intervalo de Credibilidad del 95% de la prevalencia de muerte en nuestra base de datos, la decisión más adecuada es también solicitar la RED NEURONAL. De nuevo, sólo cuando la prevalencia de muerte sea del 40,741% la decisión preferible es realizar la Regresión Logística.

Análisis de Sensibilidad sobre Prevalencia (bayesiano)



Muerte en UCIP estimada con el índice "PRISM":
Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana

Análisis de Sensibilidad sobre Prevalencia (bayesiano)



Muerte en UCIP estimada con el índice "PRISM":
Comparación del rendimiento diagnóstico de las predicciones realizadas con un modelo de regresión logística,
y una red neuronal artificial. Una propuesta bayesiana_____



Universitat d'Alacant
Universidad de Alicante



VII. - Discusión.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

Las dos actividades centrales de la medicina, el diagnóstico y el tratamiento, representan la correspondencia en el mundo de la investigación clínica de los que para Joseph B. Kadane y Teddy Seidenfeld¹⁵⁷ son los dos paradigmas o visiones distintas (inconmensurables, en el sentido semántico del primer Khun¹⁵⁸) en los que se basa la investigación científica contemporánea: la "ciencia para aprender" y la "ciencia para probar causalidad", respectivamente. Ambas visiones suelen coexistir simultáneamente, incluso en la mente de una misma persona.

1.- LA INVESTIGACIÓN CIENTÍFICA SOBRE TERAPIA: EXPERIMENTOS PARA "PROBAR CAUSALIDAD"

Para demostrar científicamente la efectividad de un tratamiento se hace imprescindible la realización de *experimentos para probar causalidad*. En ellos, mediante estudios de intervención escrupulosamente diseñados, se intenta evidenciar la presencia de una relación causal entre la administración de un tratamiento y un resultado de relevancia clínica (curación, muerte, etc...). La forma más desarrollada y rigurosa -el prototipo- de tales experimentos son los ensayos clínicos aleatorizados y controlados. Jan P. Vandembroucke llama¹⁵⁹ a estos estudios "investigación para evaluar" porque su objetivo es examinar rigurosamente si la evolución clínica de los pacientes se ve realmente modificada por la intervención de nuevas terapias que, conjeturando desde los resultados de los estudios observacionales, nos parecen inicialmente tan maravillosas o prometedoras.

La investigación observacional ("para explorar o descubrir" en la terminología de Vandembroucke¹⁵⁹) es imprescindible para el progreso de la ciencia. Pero debido al irresoluble "problema de la inducción", está legitimada sólo para conjeturar hipótesis verosímiles y su objetivo es únicamente exploratorio. La

practican médicos seducidos por nuevos descubrimientos (quizá serendípicos) o por ingeniosas y audaces explicaciones causales de la enfermedad o de su mecanismo fisiopatológico. La historia de la ciencia nos enseña que los descubrimientos científicos suelen ocurrir cuando se miran las mismas cosas de siempre con una nueva mirada: requieren libertad, perspicacia y un poco de suerte. Sorprendidos por el curso extraño de una enfermedad en un paciente individual, por unos resultados incoherentes de laboratorio, o por "la pinta peculiar" que toma un subgrupo particular de pacientes en el análisis que los datos, los investigadores se entusiasman con una idea y buscan datos (al principio preferentemente retrospectivos) sugerentes de que "pudiera existir algo". En cuanto encuentran coherencia entre los datos y su idea, pensando que están ante una confirmación indirecta, elaboran un manuscrito y lo envían a publicar. Nuevos investigadores, utilizando sus datos preexistentes o iniciando una recogida prospectiva de nuevas observaciones, intentan refrendar dicha idea. La buscan en un subgrupo diferente de enfermos, varían la definición y el grado de la exposición o tienen en cuenta fuentes potenciales de sesgo y factores de confusión, intentando explicar porque la nueva idea se cumple o porque es totalmente errónea. Tan pronto como tienen listos sus modelos, de nuevo envían sus resultados a ser publicados. Toda esta investigación no hace más que dotar a la idea de verosimilitud, o disminuirla drásticamente. Por eso, la investigación observacional requiere siempre de nuevos estudios que traten de resolver definitivamente la controversia.

La ciencia conoce profundamente la alta probabilidad de sesgos que tiene la investigación observacional: los clínicos presentan series de casos fuera de contexto o con multitud de co-intervenciones, y los epidemiólogos realizan múltiples análisis sobre datos persistentes recogidos con un propósito totalmente distinto. Como se suele decir en la jerga "si se exprimen adecuadamente, los datos -al final- acaban cantando" y todo parece cuadrar. Por eso los experimentadores evaluativos actúan de "control de calidad" científico: para no permitir que nuestro deleite en la exploración de nuevas ideas pueda afectar a la salud de los futuros pacientes. Encontrar explicaciones es un desafío intelectual, un juego cuyo premio

está reputado con el que Leonardo da Vinci llamó el placer más noble: "el goce de comprender". La mayoría de los científicos disfrutan con él, pero igualmente quedan muy contrariados cuando descubren algún error o violación del protocolo en un ensayo clínico. Conscientes del trascendental papel que la sociedad demanda a la ciencia, los médicos saben que los resultados de tales ensayos van a aplicarse a multitud de personas reales. Para evitar hacer daño a la gente exigen que sean ciertos o verosímilmente verdaderos, no sólo atractivas ideas interesantes.

1.1.- VERSIÓN FRECUENTISTA DE LA CIENCIA PARA PROBAR CAUSALIDAD:

Es ingente la literatura que explica los conceptos clave (Error tipo I, Error tipo II, poder del estudio, nivel p de significación estadística e intervalos de confianza) que utiliza la evasión frecuentista para diseñar y analizar los resultados de un *experimento para probar causalidad*. No los vamos a detallar aquí en profundidad, no es el objetivo de esta tesis. Sólo nos interesa resaltar como, bajo este paradigma estadístico clásico, las probabilidades de cometer ambos tipos de error, que se eligen arbitrariamente en la fase de diseño de los ensayos clínicos (antes de obtener los datos), se utilizan luego (una vez analizados los resultados) para realizar la inferencia o *comportamiento inductivo*. Son muchos y bien conocidos^{160,161} los graves defectos e inconsistencias lógicas de este modelo de inferencia inductiva. Así, la evidencia que se maneja no depende sólo de los datos obtenidos en el experimento (el llamado principio de verosimilitud¹⁶²), sino que depende de componentes claramente subjetivos, p. ej. de lo que el investigador ha planeado hacer antes de recoger los datos (*stopping rules* en el lenguaje técnico), del tamaño muestral elegido en la fase de diseño (el tamaño muestral creciente convierte los mismos resultados en estadísticamente significativos), de cómo se analizan los datos (técnicas paramétricas vs. no paramétricas), de niveles de significación fijados subjetivamente (aunque prácticamente aceptados por consenso universal), de datos no observados y que se consideran aún más

infrecuentes que los datos realmente obtenidos (recordemos que el nivel de significación $p = P[\text{datos} \geq \text{estimador} / H_0]$; esto es, se obtiene por integración de la(s) cola(s) de la distribución muestral bajo H_0 , y por tanto acumula probabilidad correspondiente a datos aún mas discrepantes con H_0 de los realmente observados en el experimento), o de intervalos de confianza que no pueden interpretarse como intervalos de probabilidad y por tanto no aceptan álgebra de probabilidades. De hecho, la mayoría de estadísticos coinciden en reconocer que la metodología frecuentista consiste en "...una síntesis incompletamente formalizada de ingredientes variados extraídos de fuentes teóricas mutuamente incompatibles"¹⁶³. A pesar de ello, la ciencia en general continúa abogando por el uso de estos métodos. No sólo porque la mejor alternativa disponible, la evasión bayesiana, realiza su inferencia basándose en un componente subjetivo más conspicuo, la distribución de probabilidad a priori, y presenta además el *handicap* de que necesita de un apoyo computacional y de software muchísimo más complejo. Existe otra razón importante de carácter mucho más práctico: la ciencia utiliza mayoritariamente la evasión frecuentista porque le provee de algo que la ciencia realmente necesita. Específicamente: los métodos estadístico frecuentistas (pero no los bayesianos) nos proveen de una medida objetiva explícita, y por ello de una forma de control, de la frecuencia con la que vamos a cometer errores en nuestra inferencia inductiva¹⁶².

Es Deborah G Mayo, que trabaja en la Universidad y en el Instituto Politécnico del Estado de Virginia (Estados Unidos de América), quien más rigurosamente ha estructurado^{29,164,165,166} una filosofía de la estadística frecuentista clásica o, como ella la denomina, "basada en los errores estadísticos". Para los bayesianos, el resultado más relevante de un *experimento para probar causalidad* es que estimemos la probabilidad (creencia racional) de que la hipótesis causal (llámemosla H) sea cierta a la luz de los datos experimentales: $P(H/\text{datos})$. Sin embargo, Mayo re-interpreta los test estadísticos frecuentistas clásicos como herramientas para inducir conocimiento desde los resultados de un *experimento para probar causalidad*, y razona que la evasión frecuentista no desea estimar la

probabilidad $P(H/\text{datos})$, sino que establece como momentáneamente cierta/falsa a H y por ello la acepta/rechaza completamente. Lo más sobresaliente del discurso de Mayo es que rescata una idea que subyace en el sentir común de la mayoría de los investigadores y científicos "de a pie": su argumento se basa en que los datos del experimento constituyen evidencia a favor de la verdad o falsedad de H en la medida en que H ha pasado o no un *examen severo* con ellos. El análisis estadístico clásico de los datos obtenidos tras realizar un *experimento para probar causalidad* constituye un examen para H . Solo si H pasa el examen, será momentáneamente aceptada.

Para Mayo, la dificultad o severidad del examen radica en la probabilidad de cometer error con el análisis frecuentista de los datos. Y de ahí la relevancia que otorga a los errores de la estadística clásica¹⁶⁷ (Error-I o "Falso positivo": Rechazar una hipótesis nula H_0 que es realmente cierta; Error-II o "Falso negativo": Aceptar una hipótesis nula H_0 que es realmente falsa). La probabilidad de cometer dichos errores ha de ser fijada por el investigador en la fase de diseño del estudio, antes de obtener los datos. Pero, según Mayo, es la probabilidad de cometer estos errores la que se utiliza después, una vez analizados los resultados, como medida de la severidad con la que las hipótesis han pasado (o no) el examen y deben ser aceptadas (o rechazadas). Siendo $\alpha = P(\text{cometer Error-I})$, $\beta = P(\text{cometer Error-II})$ y poder = $(1 - \beta) = P(\text{detectar una diferencia realmente existente o "Verdadero positivo"})$, un examen será muy severo para H_0 -y por tanto para H_1 , la hipótesis alternativa- si el análisis frecuentista de los datos experimentales se hace con un diseño experimental calibrado para tener α y β muy pequeños. En la teoría de la inducción de Mayo, se define una "Función de Severidad o Dificultad del Examen" que caracteriza a la inducción que se realiza sobre las hipótesis. El grado de severidad de los exámenes que pasan las hipótesis para ser aceptadas como ciertas corresponde con las probabilidades de error frecuentista: la severidad con la que se rechaza H_0 (acepta H_1) es α y la severidad con la que se acepta H_0 (rechaza H_1) es el poder del estudio $(1 - \beta)$. Si, a la luz de los datos obtenidos experimentalmente, cualquier hipótesis H pasa un examen difícil o

severo (α y β muy pequeños) tendremos una evidencia robusta de su veracidad. Si el examen es fácil, la evidencia que sustenta a H es débil.

Mayo resume su programa con la siguientes aseveraciones: "Un resultado experimental es una buena indicación de que el error está ausente si (el experimento se ha calibrado de tal modo que) hay una altísima probabilidad de que el error se hubiera detectado si existiera, y de hecho no ha sido detectado"¹⁶⁵. "La intuición nos dice que (...) los datos nos proveen de buena evidencia para inferir H (sólo) hasta el extremo en el que H pasa un examen severo con ellos, es decir, hasta el extremo de que H (muy probablemente) no hubiera sido capaz de pasar el examen si hubiera sido falsa". (...) "A la luz de los datos obtenidos experimentalmente, una hipótesis estadística H pasa un *examen severo* con esos datos si a) los datos son coherentes con H, y b) con altísima probabilidad, si H fuera falsa, el experimento hubiera arrojado un resultado que concordaría peor con H de lo que lo hacen los datos realmente obtenidos. Dicho de otra manera: si H fuera falsa, con muy baja probabilidad el experimento produciría un resultado que concordara igual de bien (o mejor) con H de lo que lo hacen los datos realmente obtenidos"¹⁶⁶.

El concepto de severidad de los exámenes a los que someter las hipótesis para aceptarlas ya está presente en Karl R Popper¹⁶⁸, pero se inserta claramente en los postulados de la más contemporánea de las escuelas epistemológicas norteamericanas, la llamada teoría "externalista" del conocimiento. Sus máximos exponentes son Alvin Goldman¹⁶⁹, y sobretudo Robert Nozick, para quien sabes algo si, y sólo si, *no lo hubieras creído de haber sido falso*. Las creencias de uno tienen que ser muy sensibles -en el sentido de cambiar de opinión ante la mínima provocación- a los hechos y a los cambios el mundo real (y extremadamente insensibles al ruido de fondo y a otras nimiedades) para ser consideradas conocimiento verdadero¹⁷⁰.

Es así como se entiende que, para la evasión frecuentista, la distribución aleatoria de los pacientes a los tratamientos sea un requisito imprescindible a la hora de analizar la validez de los resultados de los *experimentos para probar*

causalidad, si se quieren utilizar en su función de "investigación evaluativa". Sólo van a aceptarse (momentáneamente) la hipótesis que hayan pasado exámenes severos.

Cuando se realiza análisis estadístico frecuentista, la **aleatorización** de muestras grandes impide al autor controlar el diseño una vez que el experimento ha empezado, pues la asignación de los pacientes a los tratamientos se lleva a cabo mediante un mecanismo probabilística externo al investigador, y también protege (asintóticamente) contra la confusión generada por otros factores no controlados, aparte del que se está testando causalmente. Pero, además, tal y como acabamos de ver, la inferencia frecuentista se basa en la dificultad o severidad de los exámenes a los que se somete a las hipótesis. Los grados de severidad corresponden con las probabilidades de error estadístico, y éstas sólo pueden calcularse con exactitud, en base al Teorema del Límite Central, si se adopta como modelo estadístico a una distribución probabilística muestral particular (p. ej. la normal gaussiana) de todos los posibles resultados experimentales bajo la "hipótesis nula". Para que ello se cumpla con rigor es imprescindible cumplir los requisitos de aplicación de dicho Teorema, a saber: un tamaño muestral suficientemente grande y un **muestreo aleatorio**. Como a veces, sobretodo por razones logísticas, el primer requisito no se cumple, es casi obligatorio cumplir el segundo. Sólo así tiene estrictamente sentido matemático calcular los niveles de significación o los Intervalos de Confianza, y usar la estadística frecuentista clásica en la argumentación deductiva que subyace al método de inferencia de los **experimentos para probar causalidad**.

1.2.- VERSIÓN BAYESIANA DE LA CIENCIA PARA PROBAR CAUSALIDAD:

Desde el punto de vista de la evasión bayesiana, un **experimento para probar causalidad** constituye lo que la Teoría Matemática de la Decisión denomina un problema con más de un decisor. De hecho, existen -en un contexto finito- al

menos dos actores implicados, cada uno con sus propias creencias (racionales) en la veracidad de la hipótesis que se quiere probar: el autor (decisor bayesiano A) y el lector (decisor bayesiano B). El método correcto de solucionar un problema de esa naturaleza se basa en los mismos principios que rigen un análisis de decisión clásico -con un solo decisor-, pero es algo más complicado. El lector debe considerar al autor como lo que es: un decisor bayesiano que a) ha estudiado el tema a conciencia y conoce el fenómeno en profundidad; esto es: es capaz de definir una *probabilidad previa* de la veracidad de su hipótesis, y b) cuya *función de utilidad* refleja su deseo de comprobar que su hipótesis es cierta. Y por ello el lector debe saber que, en virtud del teorema de Bayes, el autor (decisor A) es capaz de actualizar su creencia (racional) en la veracidad de la hipótesis mediante los datos muestrales del experimento, para expresarla en forma de probabilidad a posteriori. Utilizando ésta y su propia *función de utilidad*, que -a diferencia de lo que ocurre con el autor- reflejará probablemente su deseo de beneficiar a los pacientes, el lector (decisor B) debería ser capaz de tomar con coherencia sus decisiones, maximizando su utilidad esperada.

Pero, ¿cómo puede el lector estar seguro de que los datos muestrales que aporta el autor son *intercambiables* en el sentido del teorema de representación? ¿Cómo puede asegurar el autor que no ha sido él quien ha construido la aparente superioridad del tratamiento experimental mediante la forma como ha asignado los pacientes a los tratamientos? Dejar tal asignación al investigador (que suele ser además el más interesado en que "parezca que" el nuevo tratamiento funcione) parece peligroso, y por ello el diseño experimental elegido por el autor constituye, para el lector, la principal fuente legítima de preocupación.

Stone primero^{171,172}, pero sobretudo Rubin^{104,173}, han subrayado que la asignación aleatoria de los pacientes a los tratamientos es la forma más sencilla que tiene el investigador de asegurar al lector la intercambiabilidad de sus datos muestrales, y justificar el uso de las probabilidades a priori. Al igual que ocurre con los frecuentistas, para los bayesianos la *aleatorización* de muestras grandes impide al autor controlar el diseño, y protege (asintóticamente) contra la

confusión. Pero en este caso posibilita, en base al Teorema de Representación, la adopción de un modelo estadístico particular bajo el que analizar los resultados experimentales y poder obtener la distribución de probabilidad a posteriori (modelo Beta-Binomial, modelo Normal-Student, modelo Gamma-Poisson, etc...). Y además facilita mucho la labor del lector (decisor B), por dos motivos fundamentales. Primero, porque el uso de cualquier otro diseño comparable pero no aleatorizado resultaría sustancialmente muchísimo más difícil de analizar, debido a la necesidad de asegurar que se ha modelizado el mecanismo de muestreo y asignación de los tratamientos (en los estudios observacionales el mecanismo de asignación de tratamientos suele ser desconocido) y que todas las covariables han sido controladas. Y segundo, porque sería necesario que el autor (decisor A) hiciera matemáticamente explícita su distribución *a priori* para que el lector (decisor B) pudiera hacer uso exclusivamente de la información muestral vehiculizada por la función de verosimilitud, para llevar a cabo rigurosamente su propio análisis de decisión. La aleatorización permite al lector (decisor B) ignorar las distribuciones *a priori* y las funciones de utilidad del investigador (decisor A), y simplifica la comunicación científica entre ellos. Es una sabia elección que un autor intente facilitar las cosas al lector evitándole todos esos esfuerzos, incentivando así la lectura de sus originales.

Así pues, también desde el punto de vista de la evasión bayesiana, la aleatorización de muestras grandes desempeña una importantísima función para asegurar la validez científica de los *experimentos para probar causalidad* (investigación evaluativo).

2.- LA INVESTIGACIÓN CIENTÍFICA SOBRE DIAGNÓSTICO: EXPERIMENTOS PARA "APRENDER"

Sin embargo, la actividad diagnóstica de la práctica clínica se encuadra mejor en el paradigma que Kadane y Seidenfeld¹⁷⁴ denominan "ciencia para aprender". Cada vez que un médico se enfrenta a un paciente en una situación de incertidumbre sobre la presencia o no de una enfermedad particular, realiza uno de los llamados *experimentos para aprender*. La forma científicamente más correcta de llevar a cabo uno de estos experimentos es mediante lo que matemáticamente se denomina Problema de Decisión Clásico, con un solo decisor implicado. Para realizar un diagnóstico a ese enfermo concreto, es obvio que el médico no puede basarse en la repetición un número elevado de veces de esa misma situación con esa misma persona, para poder obtener una conclusión frecuentista. El enfermo no desea que el médico acierte asintóticamente con el 95% de los pacientes que son semejantes a él, sino que exige del médico que tome (a veces incluso rápidamente) decisiones terapéuticas precisamente con él: una persona particular en una situación determinada, pero en un contexto de incertidumbre. Por ello el médico, cuando realiza su actividad diagnóstica, sólo puede hacer uso de la evasión bayesiana al problema de la inducción, para la que el concepto de probabilidad no se relaciona con una estabilidad (asintótica) de las frecuencias con la repetición de eventos, sino con la creencia (racional) en la presencia de una enfermedad en un enfermo determinado en ese momento concreto del tiempo. Mediante el uso del Teorema de Bayes, el médico realizará un *experimento para aprender*, y será capaz de actualizar esta creencia en la presencia de enfermedad a la luz de los datos obtenidos de la anamnesis, la exploración física y los resultados de los test diagnósticos. Esa creencia actualizada en forma de distribución a posteriori de probabilidad, puede ya incorporarse en un Análisis de Decisión para poner en

marcha coherentemente las medidas terapéuticas más adecuadas a cada situación particular.

Cuando se hace uso de la evasión bayesiana, y el objetivo es aprender de los datos de experimentación para actualizar nuestra creencia (racional), es de hecho innecesario y probablemente subóptimo realizar muestreo aleatorio: no se van a obtener conclusiones basadas en las distribuciones muestrales ni en el Teorema del Límite Central. Se van a obtener, mediante el Teorema de Bayes, distribuciones a posteriori de probabilidad, y para ello sólo se va a exigir un juicio (subjetivo) de *intercambiabilidad* de los datos que permita -en base al Teorema de Representación- generar las distribuciones a priori. Se necesita sólo que el médico que va a hacer el diagnóstico juzgue que el paciente que tiene delante es intercambiable con los pacientes con los que se han calculado los índices de exactitud diagnóstica de los test que va a utilizar para intentar confirmar o descartar la presencia de enfermedad.

Y aquí es donde radica la paradoja que ha dado origen a esta tesis doctoral. Los organismos internacionales (iniciativa STARD) que definen los *standards* para el diseño, realización y publicación en la literatura médica contemporánea de estudios que evalúen la exactitud de los test diagnósticos¹⁷⁵, sin exigir que la selección de los pacientes de la muestra se realice al azar, recomiendan la utilización de métodos estadísticos frecuentistas para el cálculo de los Intervalos de Confianza^{176,177} de los índices que miden el rendimiento de dichos tests diagnósticos (sensibilidad, especificidad, razones de verosimilitud, curvas ROC, etc...). Nos parece que ello es un contrasentido matemático y un grave error metodológico. Este estudio demuestra que es posible utilizar metodología estrictamente bayesiana para estimar la magnitud y la precisión de los índices de exactitud diagnóstica de una prueba, con sus respectivos Intervalos de Credibilidad.

2.1.- ESTUDIO DE LA CAPACIDAD DISCRIMINANTE:

Desde el punto de vista frecuentista, en la validación externa el comportamiento discriminante de ambos modelos fue excelente. Al estudiar la curva ROC, el AUC del modelo de RL fue de 0,9412 ($p < 0,000$), y el de la RN fue de 0,9552 ($p < 0,000$). Unos intervalos de confianza para estas medidas, según los métodos utilizados, quedan recogidos en la siguiente tabla:

	AUC	Intervalo de confianza asintótico al 95%		Método
		Límite inferior	Límite superior	
RL	0,9412	0,8774	1,0049	DeLong
		0,8336	1,0488	Hanley & McNeil
RN	0,9552	0,9102	1,0002	DeLong
		0,8602	1,0501	Hanley & McNeil

Bamber introdujo el uso de la curva ROC (del inglés *Receiver* o *Response Operating Characteristic*) en ciencias de la salud¹⁷⁸, un instrumento tomado de la teoría de detección de señales que se usa en ingeniería, y que se convirtió en una ayuda crucial para distinguir las señales reales de los ruidos falsos en los inicios del radar (segunda guerra mundial). Muy extensa es la literatura médica sobre estimación de área bajo la curva ROC^{179,180}, su Intervalo de Confianza, y su significación estadística. Pero los métodos frecuentistas clásicos, no sólo no tienen sentido matemático si no ha existido muestreo aleatorio, sino que su uso no está exento de problemas: no funcionan con tamaños muestrales pequeños¹⁸¹, no pueden calcularse para tests con sensibilidades o porcentajes de falsos positivos cercanas a 0 o a 100%, y a veces extienden su intervalo de confianza por fuera del gráfico de la curva¹⁸². Este problema es exactamente lo que ocurre en este trabajo. Como puede apreciarse claramente en la tabla anterior, tanto utilizando el método de DeLong como el de Hanley-McNeil, la estimación frecuentista del IC95% del AUC

en ambos métodos diagnósticos nos dá unos límites superiores mayores de 1. Si consideramos -por definición- que el AUC es una probabilidad, ciertamente estos valores no tienen ningún sentido: una de las características definitorias de una probabilidad es que su valor es siempre inferior o igual a 1.

Estudios de simulación han mostrado que, cuando se utilizan muestras pequeñas, los IC95% frecuentistas del AUC de la curva ROC tienen una mala propiedad de "cobertura" (ver más adelante), y por ello se considera que no existe aún una única alternativa óptima para construir tales ICs¹⁸³. Por otra parte, es ciertamente escasa la bibliografía que describe métodos bayesianos para la construcción de curvas ROC. Además resultan tremendamente complicados y poco accesibles a médicos no familiarizados con el uso de técnicas matemáticas superiores. Así Peng y Hall¹⁸⁴, mejorando el enfoque previo de Tosteson y Beg¹⁸⁵, ajustan la curva ROC a una modelo de regresión bi-normal, y proponen métodos de cadenas de Markov y Monte Carlo (muestreo de Gibbs) para estimar los parámetros del modelo y luego dibujar una curva ROC circunscrita en el interior de una región de credibilidad, y estimar el área que queda bajo ambos límites. En un artículo posterior, Hellmich y colaboradores¹⁸⁶ generalizan el método para un modelo de regresión general no lineal, y lo aplican mediante muestreo de Gibbs a un ejemplo particular. Recientemente, Tibury y colaboradores¹⁸² han desarrollado un programa de ordenador en C++ que estima el contorno de la curva ROC con el 95% de credibilidad bajo la distribución posterior, y lo han comprobado en un estudio de simulación de Monte Carlo.

Al estudiar frecuentísticamente los índices de exactitud diagnóstica hemos obtenido unos resultados desconcertantes. Dada la existencia en la tabla de 2x2 de una casilla con ceros, en el modelo de RL la Especificidad ha sido de 76% pero la Sensibilidad ha sido de 100% con un IC 95% = 100% a 100%, un valor sin sentido para una intervalo de Confianza por su sentido determinista. Con estos valores, ha sido imposible calcular la RV para negativos (no se puede dividir por cero), y el valor de la RV para positivos es un discreto 4,17 (IC95% = 3,19 a 5,45). Los índices de exactitud han sido un poco mejores en la RN: la Especificidad ha sido del 87,1%,

pero la Sensibilidad ha sido del 88,9% con un IC95% = 68,4% a 109,4%. Como puede comprenderse este límite superior es un valor sin sentido: la Sensibilidad es una probabilidad y por definición su valor no puede ser superior a 1 (o a 100% si se expresa en porcentaje). Sí hemos podido calcular las RV de la RN: RV para positivos 6,91 (IC 95% = 4,39 a 10,87) y la RV para negativos 0,13 (IC 95% = 0,02 a 0,81).

Con todo ello, desde el punto de vista de la comparación frecuentista no hemos podido descartar la hipótesis nula de igual capacidad discriminante entre los dos modelos evaluados: Ji-Cuadrado de DeLong = 0,8862 con grados de libertad = 1, $p = 0,3465$. No se pueden estimar Intervalos de Confianza de la diferencia entre AUCs, pues Ji-cuadrado es un estadígrafo no paramétrico.

Desde el punto de vista de nuestra propuesta bayesiana, también ambos modelos tuvieron una excelente capacidad discriminante en la validación de la cohorte externa. La aproximación bayesiana que presentamos en este trabajo es mucho más sencilla, y parte de la definición del área bajo la curva ROC como una probabilidad¹⁷⁸, por lo que se aplica simplemente la metodología bayesiana desarrollada para estimar probabilidades: el análisis de referencia utilizando la distribución Beta no informativa y su conjugada, la verosimilitud Binomial. Las matemáticas son muy simples, y la mayoría de los cálculos se pueden realizar con una sencilla hoja de cálculo en cualquier computadora personal. Del estudio bayesiano de las curvas ROC se obtuvo que para el modelo de RL se puede estimar una probabilidad del 95% de que el AUC se encuentre entre 0,9 y 0,969, siendo el valor más probable $AUC = 0,9365$, y para el modelo de RN se puede estimar una probabilidad del 95% de que el AUC se encuentre entre 0,922 y 0,981, siendo el valor más probable 0,953. En ambos casos estamos seguros, prácticamente al 100%, de que la capacidad discriminante de ambos modelos es superior al 0,5 (el poder de una moneda al azar). Todos los valores estimados de AUC fueron números entre 0 y 1, como corresponde al concepto de probabilidad.

El mismo método sencillo se ha empleado para estimar la versión bayesiana de los principales índices de exactitud diagnóstica. Sensibilidad y Especificidad

son probabilidades, y sus Intervalos de Credibilidad se han calculado con el modelo conjugado Beta-Binomial partiendo de distribuciones a priori de referencia no informativas (Beta[1/2, 1/2]). Desde el punto de vista de la epidemiología, los índices que tienen más valor para la toma de decisiones clínicas son las Razones de Verosimilitud, cuyo valor oscila entre 0 e ∞ . Su logaritmo neperiano (LnRV), una variable cuyo recorrido se extiende desde $-\infty$ a ∞ , tiene una larga tradición en la historia de la estadística: se denomina "peso de la evidencia"⁷⁴, "función de soporte"¹⁸⁷ o "soporte evidencial que aportan los datos"¹⁸⁸, y fue la que utilizó Alan Turing para descubrir el código enigma de los nazis. Se han desarrollado métodos frecuentistas clásicos para aproximar el cálculo de los límites del Intervalo de Confianza de una Razón de Verosimilitudes¹⁸⁹. Para un enfoque bayesiano general, el cálculo de las Razones de Verosimilitud precisa de la integración de las funciones de verosimilitud. Nuestra propuesta para calcular sus Intervalos de Credibilidad ha consistido en un muestreo de Gibbs utilizando cadenas de Markov y métodos de Monte Carlo mediante el programa WinBUGS. El uso del programa es muy sencillo, y el resultado tremendamente satisfactorio.

En el estudio bayesiano de ambos modelos pudieron computarse todos los índices de exactitud diagnóstica, y sus valores siempre tuvieron sentido matemático. La RL presentó una Sensibilidad del 95%, con una probabilidad del 95% de que su valor se encuentre entre 76,24 y 99,999%, y una Especificidad del 76%, con una probabilidad del 95% de que su valor se encuentre entre el 69,4 y el 82,1%. La RN mostró una Sensibilidad del 85%, con una probabilidad del 95% de que su valor se encuentre entre 62,3 y 99,8%, y una Especificidad del 87%, con una probabilidad del 95% de que su valor oscile entre 81,8 y 91,7%.

Pero donde la aproximación bayesiana ha demostrado más su utilidad es en la estimación de las RVs. El modelo de RL mostró una mediana de RV para positivos de 4,011, con una probabilidad del 95% de que su verdadero valor se encuentre entre 2,477 y 5,54, y una mediana de RV para negativos de 0,032, con una probabilidad del 95% de que su verdadero valor quede entre 0,00007 y 0,3156. Por su parte, el modelo de RN mostró una mediana de RV para positivos de 6,752, con

una probabilidad del 95% de que su verdadero valor se encuentre entre 3,985 y 10,47, y una mediana de RV para negativos de 0,1731, con una probabilidad del 95% de que su verdadero valor quede entre 0,01391 y 0,4796. Con la aproximación bayesiana podemos además calcular: a) la probabilidad de que el modelo de RL sea un buen método diagnóstico para confirmar la muerte es bajísima [$P(RV_{pos} > 5/RL) = 0,0365$], y sin embargo la probabilidad de que se comporte como un buen método diagnóstico para descartar la muerte es altísima [$P(RV_{neg} < 1/5/RL) = 0,9183$]; y b) la probabilidad de que el modelo de RN sea un buen método diagnóstico para confirmar la muerte es muy alta [$P(RV_{pos} > 5/RN) = 0,8738$], y la probabilidad de que se comporte como un buen método diagnóstico para descartar la muerte es bastante aceptable [$P(RV_{neg} < 1/5/RN) = 0,655$]. Estos cálculos ayudan mucho desde el punto de vista de su uso clínico, porque nos informan sobre la capacidad diagnóstica de ambos modelos, pero es imposible realizarlos con la aproximación frecuentista clásica.

Nuestra propuesta de comparación bayesiana de la capacidad discriminante de ambos modelos tiene tres modalidades de análisis, por lo que puede obtenerse una información distinta y mucho más completa de la que se obtiene con la comparación clásica. Para las tres, la hipótesis más verosímil es la de que el modelo de RN es más discriminante que el de RL, una conclusión muy importante que no se puede alcanzar usando el análisis frecuentista.

Utilizando la *aproximación normal* podemos estimar que la distribución posterior sobre las diferencias en el AUC bajo la curva ROC de los modelos de RN y RL es la Normal (media = - 0,0166, desv típ = 0,0239). Con ello, podemos calcular que tenemos una probabilidad del 95% de que la verdadera diferencia entre las AUCs se encuentre entre - 0,0635 y 0,0303. Pero además, por integración, podemos saber que la probabilidad de que la capacidad discriminante de la RL sea superior (esto es, la probabilidad de que el AUC sea mayor en la RL que en la RN) es de 0,2437, y la probabilidad de que la capacidad discriminante de la RN sea superior (la probabilidad de que el AUC de la RN sea mayor) es 0,7563. La hipótesis de que la RN sea un modelo más discriminante que la RL es 3 veces más

probable que la hipótesis contraria: se trata de una razón de estas dos verosimilitudes.

$$\frac{P(RN_{mejor})}{P(RL_{mejor})} = \frac{P(AUC_{RN} > AUC_{RL})}{P(AUC_{RL} > AUC_{RN})} = 0,7563 / 0,2437 = 3,1$$

Utilizando la *simulación de Monte Carlo* podemos obtener con el programa EpiDat 3.1 la distribución posterior sobre la distribución de la diferencia entre las AUCs, y estimar sus percentiles más relevantes (mediana P50 = - 0,016; P2,5 = - 0,066; P97,5 = 0,030): el cero es un valor aceptable. Pero nos permite también estimar estas probabilidades. La probabilidad de que la capacidad discriminante sea superior en el modelo de RL es ahora de 0,242 y la de que la capacidad discriminante sea superior en la RN es de 0,758. De nuevo la hipótesis de superioridad de la RN es tres veces más verosímil:

$$\frac{P(RN_{mejor})}{P(RL_{mejor})} = \frac{P(AUC_{RN} > AUC_{RL})}{P(AUC_{RL} > AUC_{RN})} = 0,758 / 0,242 = 3,13$$

El *método de la parrilla de Berry* es mucho más "burdo", porque se trata de una aproximación discreta al problema, aunque para su aplicación sólo se necesita una hoja de cálculo disponible en la inmensa mayoría de las computadoras de sobremesa. El gráfico nos indica que la diagonal está incluida en la "pirámide", con lo que de nuevo puede concluirse que el cero es un valor aceptable de la diferencia entre las AUCs, pero puede estimarse de nuevo que la hipótesis de superioridad del modelo de RN es más verosímil que su contraria. En este caso el cálculo de la razón de las verosimilitudes entre las dos hipótesis arroja una cifra superior 8, a nuestro parecer una estimación exagerada:

$$\frac{P(RN_{mejor})}{P(RL_{mejor})} = \frac{P(\text{difAUC}(RN - RL) > 0)}{P(\text{difAUC}(RN - RL) < 0)} = 0,8906 / 0,1094 = 8,19$$

Como puede observarse, calculados mediante un análisis de referencia no informativo, los límites de los Intervalos de Credibilidad (bayesianos) y de Confianza (frecuentistas) prácticamente coinciden. Es de destacar que los bayesianos, que se obtienen al integrar la distribución a posteriori, son generalmente más estrechos, lo que aumenta su precisión: son más informativos. Y

su interpretación conceptual es directa, y más adecuada a la mentalidad que subyace al proceso de diagnóstico¹⁹⁰ médico.

La coincidencia numérica entre los Intervalos de Confianza clásicos y los bayesianos Intervalos de Credibilidad, cuando estos se han calculado utilizando distribuciones a priori planas o no-informativas, es un fenómeno bien estudiado en la literatura especializada^{191,203} que ocurre con mucha frecuencia en la práctica. Y es por ello que, para muchísimos autores, la ideas bayesianas no son más que nimiedades filosóficas irrelevantes. En este sentido nuestra preferencia coincide con la de autores como Zech¹⁹² y D'Agostini¹⁶⁰, y es que las cosas deberían verse exactamente desde el otro punto de vista. En rigor, los Intervalos de Confianza frecuentistas no tienen ningún sentido, a no ser que coincidan con los Intervalos de Credibilidad bayesianos calculados bajo condiciones bien definidas. Nada en la estadística clásica parece tener ningún significado inmediato para nuestra forma natural de pensar, a no ser que, con nuestro sentido común, lo dotemos de una interpretación bayesiana. La mayoría de los médicos y de los investigadores clínicos están intelectualmente a favor de las definiciones frecuentistas, pero en la práctica utilizan las bayesianas. Y, de hecho, en sus mentes términos como Intervalo de Confianza o valor p de significación estadística sólo pueden vivir parasitando ilícitamente un uso bayesiano de los conceptos.

Para explicar nuestra posición, reproduciremos un diálogo ficticio basado en el inventado por el estadístico canadiense George Gabor (publicado en el libro de D'Agostini¹⁶⁰) y en un famoso artículo¹⁹³ para físicos. Retomaremos así la manera socrática de argumentar mediante el diálogo, una forma de discurso que se ha utilizado con bastante repercusión en la literatura contemporánea de nuestro país^{194,195,196} sobre el debate frecuentista vs. bayesiano. El texto simula reproducir el intercambio de opiniones entre un alumno incisivo y difícil de convencer (**A**) y su profesor (**P**), en un típico curso de introducción a la estadística. La clase está justo en el momento en el que, por primera vez, se ha introducido el concepto de Intervalo de Confianza (IC) clásico para una media normal, y se ha ilustrado con un ejemplo:

- P. ... y así un IC 95% para la media μ desconocida es $IC_{95\%} \mu = 1.23$ a 2.34 .
- A. (Levantando la mano) Perdóneme, señor. Justo hace sólo unos minutos usted ha enfatizado que el significado más común de un IC es en términos del concepto clásico original de "cobertura", el cual se sigue del método mediante el cual se ha construido. Nos ha explicado, que este concepto se establece en términos de un hipotético conjunto de experimentos similares cada uno de los cuales mide una media muestral m y computa un IC para μ , por ejemplo al 95% de confianza. Entonces, la construcción clásica garantiza que, en el límite de un gran conjunto, el 95% de los ICs así construidos contendrán al valor real desconocido μ , es decir "cubrirán" a μ .
- P. Así es, en efecto. Esta propiedad, denominada "cobertura" en el sentido frecuentista, es la propiedad definitoria de los ICs clásicos. Es importante entender esta propiedad como lo que es: refleja la frecuencia (porcentaje) con la que la afirmación " μ está entre 1.23 a 2.34" es cierta: el 95% de las veces. Las variables probabilísticas de esta afirmación son el límite superior (LS) y el límite inferior (LI), que en este caso particular han tomado valores 1.23 y 2.34, respectivamente. Pero μ es un valor fijo, aunque desconocido.
- A. De acuerdo, pero...
- P. Es igualmente importante entender lo que no es la "cobertura" frecuentista. No es una afirmación sobre el grado de creencia (probabilidad) de que μ esté contenida en el IC de un experimento particular.
- A. Ah, ¿No?
- P. No. El concepto de "grado de creencia" no existe con respecto a los IC clásicos, que están inteligentemente definidos mediante una construcción que se ciñe estrictamente a afirmaciones sobre $P(m/\mu)$ y evita deliberadamente utilizar una densidad de probabilidad sobre μ . Nunca considera a μ como una variable aleatoria.
- A. Pero entonces un IC es una especie de intervalo aleatorio con ciertas propiedades de cobertura, siempre que se REPITA este experimento en condiciones casi idénticas.
- P. Correcto. Pero también los IC clásicos poseen otra propiedad mucho más poderosa: si en un conjunto de experimentos reales y diferentes, cada uno de ellos mide cualesquiera de las características observables que desee y construye sobre ellas un IC, por ejemplo al 95%, entonces, en el largo plazo, el 95% de los IC así construidos cubrirán el verdadero valor de sus respectivas características. Esto tiene, como veis, una aplicación directa a la vida real, y en ello radica la belleza de los ICs frecuentistas clásicos. ¿No os parece?
- A. Si, profesor. Parece que la "cobertura" es el objetivo mágico de los IC clásicos. Supone una propiedad tal vez atractiva desde un punto de vista puramente estético, pero en realidad es también una especie de principio democrático que otorga a diferentes valores posibles del verdadero parámetro la misma probabilidad de ser incluidos en un IC. Así que no es muy obvio como podemos hacer un uso natural de este concepto. Porque... ¿Cuál es, por ejemplo, el significado real de ESTE IC concreto que acabamos de calcular?
- P. Bueno, es una de las muchísimas posibles realizaciones de una colección de intervalos. Un miembro de un conjunto, el cual tiene ciertas propiedades.
- A. Ya, pero las propiedades son del conjunto, no las posee cada elemento. ¿O es que las propiedades pueden atribuirse a cada uno de sus miembros?
- P. No, sería peor que incorrecto achacar a cada IC las propiedades del conjunto. Su interpretación probabilística pertenece sólo al conjunto.

- A. Pero entonces nuestro IC no tiene ningún significado directo.
- P. Bueno.... No exactamente. De hecho existe una manera, llamada estadística bayesiana, de atribuir un significado probabilístico simple a un IC concreto, computado en un experimento único. Pero su método de cálculo excede los objetivos de este curso básico. De todas formas, puedo asegurarnos que no existe diferencia numérica alguna entre el resultado de las dos maneras de construirlos. Si alguien tiene interés, al final de la clase....
- A. (Inquieto) ¿Quiere decir que siempre coinciden?
- P. No. Para que coincidan, deben calcularse asumiendo que no tenemos ninguna razón, antes de obtener los datos, para creer que la media desconocida μ se encuentra en algún estrecho segmento particular de la recta real.
- A. Vale, de acuerdo. Pero, ¿me permite otra pregunta? He notado que, cuando nos lo presentaba, usted lo ha llamado "un" IC, y no "el" IC. ¿Es que existen otros?
- P. Si. De hecho, existen multitud de maneras de obtener ICs, todos en conjunto con las mismas propiedades de cobertura. Sólo el de la pizarra es un intervalo bayesiano (por supuesto, si se ha calculado como os he dicho).
- A. ¿Es el bayesianismo la única manera de justificar el uso de este IC particular?
- P. No. Existen otras maneras, pero son muy complicadas. Por ejemplo, el Intervalo de Verosimilitud, una cosa extraña que no opera con probabilidad...

El diálogo podría continuarse hasta el infinito, y podríamos extenderlo para que tratara prácticamente cada uno de los conceptos con los que opera la estadística frecuentista: distribuciones muestrales de los estimadores, el concepto mismo de variable independiente e idénticamente distribuida, la significación estadística, etc.... Lo bien cierto es que, en la inmensa mayoría de los estudios de investigación clínica, por inercia, tradición, o -simplemente- porque "nos funciona", se sigue utilizando la metodología frecuentista, incluso en la literatura sobre pruebas diagnósticas.

Los principales editores de revistas médicas¹⁹⁷ y las recomendaciones estadísticas de las autoridades reguladoras internacionales¹⁹⁸ mencionan explícitamente la posibilidad de realizar análisis bayesianos. En el caso de los *experimentos para aprender*, lo único que se puede decir a favor de los métodos frecuentistas es que funcionan siempre y cuando no se viole el principio de verosimilitud. Si eso pasa (y ocurre con frecuencia), numéricamente coinciden con el resultado obtenido mediante metodología bayesiana con distribuciones a priori planas.

Nos parece que la preponderancia que otorgan muchos médicos (e incluso epidemiólogos) a los IC clásicos, obtenidos con metodología frecuentista, obedece

en realidad a una ilusión. De manera inadvertida, suelen mezclar el concepto de "cobertura" (tal y como se ha introducido en el anterior diálogo) con el de la probabilidad de que el verdadero valor del parámetro quede localizado en el interior de los límites del intervalo, incluso cuando intelectualmente comprenden que existe una clara diferencia. De hecho, la única manera matemáticamente posible de calcular tal probabilidad consiste en inventar una distribución bayesiana a priori y multiplicarla por la función de verosimilitud e integrar ese producto sobre los límites de dicho IC. Como ya se ha dicho, la "cobertura" es una propiedad de un conjunto de experimentos, realizados todos bajo unas determinadas condiciones (p. ej aleatorización de la muestra), y sólo tiene sentido si existe repetición del experimento en las mismas condiciones. Incluso los frequentistas admiten que los límites de los IC clásicos no son adecuados para tomar decisiones. El diagnóstico médico es justamente un escenario en el que se toman decisiones importantes, y no es posible la repetición muestral ni la aleatorización. Para un experimento simple y único, no podemos deducir de la "cobertura" calculada una probabilidad de que nuestro resultado (nuestro diagnóstico) sea verdadero, y es esto último lo que realmente necesitamos para tomar correctamente nuestras decisiones.

Por todo ello en esta tesis doctoral nos hemos decidido a utilizar la metodología correcta¹⁹⁹, no su sustituta. Nos parece más adecuado, científicamente, asumir abiertamente las limitaciones de la metodología bayesiana de referencia ("Dadas estas distribuciones a priori -planas o no informativas-, estos modelos probabilísticos y estos datos empíricos, creemos estar al 95% seguros de que el parámetro está incluido en este intervalo"), que no especificar en nuestro trabajo lo que, en honor a la verdad, deberíamos decir. Cuando no existe aleatorización y hacemos inferencia mediante estadística frequentista, si obtuviéramos un resultado estadísticamente significativo, lo honesto sería concluir²⁰⁰: "Si estos datos hubieran sido generados mediante un muestreo aleatorio (o en un ensayo con distribución aleatoria) sin datos perdidos ni errores de medida, el conjunto formado por estos resultados obtenidos y todo el resto de los resultados más

extremos es súmanente improbable si la hipótesis nula fuera cierta. Pero como los datos no han sido así generados, en realidad podemos decir muy poco de su verdadera significación estadística". Aunque²⁰¹, "de lo que no se puede hablar, mejor es callar la boca".

De todas maneras, y para los aún incrédulos, varios autores han estudiado empíricamente, mediante técnicas de simulación de Monte Carlo, las propiedades de "cobertura" que exhiben los Intervalos de Credibilidad bayesianos. Para el modelo binomial, los resultados muestran que si se utiliza la distribución a priori de referencia no informativa Beta (1/2, 1/2) -que es la que nosotros hemos usado-, los intervalos bayesianos obtenidos con el Teorema de Bayes tienen una cobertura excelente: tienen una probabilidad posterior casi idéntica al nivel de confianza deseado^{202,203}. De hecho, se ha establecido²⁰⁴ que los métodos bayesianos objetivos son el camino más prometedor para conseguir la unificación de las dos evasiones, la frecuentista y la bayesiana.

2.2.- ESTUDIO DE LA CALIBRACIÓN:

Debido probablemente a su gran disponibilidad en los paquetes estadísticos de software y a la facilidad de interpretación de sus resultados, la manera hegemónica en los estudios clínicos de estudiar la calibración desde el punto de vista frecuentista es con el test "bondad de ajuste" ji-cuadrado de Hosmer-Lemeshow. A diferencia de los test de hipótesis más comunes, en este caso lo que se desea es que no haya significación estadística, pues la prueba se propone para enjuiciar que la hipótesis nula (que en este caso afirma que cierta distribución rige para nuestros datos muestrales) es válida en determinado contexto. De nuevo esta prueba está basada en un test de significación estadística, y por ello tiene sus mismas deficiencias. El objetivo con el que se calcula la significación estadística en este caso es validar un modelo, pero es bien sabido que, aunque muchos de ellos resultan útiles, todos los modelos son incorrectos. En tanto que representaciones *idealizadas* de la realidad, los modelos son imperfectos por definición. Por ello

sorprende sobremanera el uso de las pruebas de significación estadística en este contexto: consideran útil un modelo por el sólo hecho de que no se ha podido demostrar que es imperfecto (cuando de hecho ya se sabe que lo es). Si la hipótesis nula afirma -como ocurre en esta prueba de "bondad de ajuste"- que los datos muestrales siguen exactamente cierta distribución, entonces *sensu stricto* esta hipótesis *siempre* es falsa. Por lo tanto, dicha hipótesis será rechazada inexorablemente si la muestra es suficientemente grande²⁰⁵. La mejor manera de conseguir lo que deseamos con esta prueba de "bondad de ajuste" es adoptar la absurda medida cautelar de no tomar una muestra demasiado grande.

Además, como ha señalado Tsiatis²⁰⁶, con mucha frecuencia se dá la situación en que el número de observados y/o esperados de cada categoría de riesgo en las que se ha dividido la muestra es menor de 5 individuos, y en estas condiciones, el estadígrafo de Hosmer-Lemeshow no se distribuye Ji-cuadrado, y lamentablemente el test no puede computarse²⁰⁷. También se ha comprobado, como ocurre con todas las pruebas frecuentistas de significación estadística, que el resultado es muy dependiente del tamaño muestral de los grupos de riesgo²⁰⁸ y de la forma como éstos se organicen.

Se han propuesto varias soluciones frecuentistas a todos estos problemas, desde la aplicación de pruebas de bondad de ajuste más complicadas²⁰⁹ (como la prueba de Farrington²¹⁰) a la estrategia de cotejar directamente los valores observados y esperados mediante simple inspección visual y evaluar el grado de concordancia a partir del sentido común²⁰⁷, pero el test de Hosmer-Lemeshow se sigue utilizando de manera casi universal. En la cohorte de validación de nuestro estudio, no hemos podido descartar la hipótesis nula ni con el modelo de RL (χ^2 de Hosmer-Lemeshow = 1,2472; grados libertad = 7; p = 0,9898) ni con el de RN (χ^2 de Hosmer-Lemeshow = 0,8585; grados libertad = 7; p = 0,9968), por lo que desde el punto de vista frecuentista asumimos que ambos están aceptablemente calibrados. Tal y como nos ha ocurrido con la capacidad discriminante, desde el punto de vista frecuentista no hemos podido estimar cual de los dos modelos presenta mejor calibración.

Desde el punto de vista bayesiano, las pruebas de bondad del ajuste han sido también muy criticadas. En virtud del principio de verosimilitud no se pueden calcular, ya que no existe un parámetro diferenciador entre la hipótesis de la que se trata y todo el conjunto "difuso" de las posibles alternativas. Por eso la mayoría de los textos aconsejan utilizar las alternativas frecuentistas (a pesar de los problemas que, como se ha visto, entrañan). Nuestra propuesta en esta tesis, es un poco más ambiciosa, y tal vez arriesgada.

En 1948 Claude Shannon, ingeniero y matemático fallecido en 2001, inició el desarrollo de la llamada teoría matemática de la información²¹¹. Se basa en un concepto algo oscuro²¹², el de entropía (información esperada) como medida del grado de indeterminación o aleatoriedad de un sistema, pero ha posibilitado toda la revolución informática y ha alcanzado en nuestros días tal preponderancia²¹³ que mucha gente llama a la nuestra la "era de la información". Sin embargo, su aplicación a la medicina ha sido casi nula. Aparte de en dos estudios de patrones electrofisiológicos de caos en pacientes con epilepsia^{214,215}, sólo hemos sido capaces de localizar en PubMed conceptos de teoría de la información en unos pocos estudios sobre evaluación de métodos diagnósticos. Benish usa gráficos de información para comparar test diagnósticos²¹⁶, y utiliza la información mutua como índice del rendimiento de un test diagnóstico²¹⁷. Okagaki y colaboradores utilizan la teoría de la información como medida de rendimiento de métodos diagnósticos en citología²¹⁸. Dragalin et al emplea la divergencia de Kullback-Leibler para evaluar bioequivalencia entre fármacos²¹⁹, y Lee la utiliza para estudiar el potencial de un test diagnóstico como descartador o confirmador de un diagnóstico²²⁰. En España, Xavier de Salvador ha utilizado la teoría de la información en estudios de modelos causales en sociología²²¹.

En este estudio, proponemos el uso de la divergencia de Kullback-Leibler^{222,223} y de la discrepancia intrínseca -que presenta propiedades matemáticas que la hacen superior²²⁴- para estimar la calibración los modelos, sin utilizar pruebas de significación estadística. El método, una modificación nuestra de la prueba de Hosmer y Lemeshow¹²⁵, pretende ocupar el lugar de este último en

la evaluación bayesiana de la calibración de un método diagnóstico que estime probabilidades. Se puede aplicar también mediante una sencilla hoja de cálculo, y su interpretación es muy intuitiva, toda vez que se mide en bits: la gente familiarizada con la informática personal entiende lo que significa actualizar, por ejemplo, su versión de Windows^R mediante un CD que contenga un programa de 4 megaBits. Así, una vez la mortalidad ha sido estimada por el modelo de RL o por la RNA, la prueba de discrepancia que proponemos nos indica el tamaño que debería tener (en bits) nuestro programa "de actualización" para obtener, a partir de la estimada, la mortalidad real. Un valor inferior a 0'01 bits indica un ajuste casi perfecto del modelo.

De nuevo en este caso, la perspectiva bayesiana por la que hemos optado nos aporta nuevos e importantes matices para poder evaluar mejor la calibración de ambos modelos. En la cohorte de validación, respecto a la distribución de las muertes observadas, tanto en el modelo de RL como en el de RN la distancia entrópica de la distribución de las muertes predichas es despreciable (véase tabla de la pág. 176, reproducida a continuación),

Distancia entrópica a lo observado

	Est. Bayesiano Convencional		Est. Bayesiano Intrínseco	
Predicción realizada por	Divergencia de Jeffreys $J(o, e)$	Discrepancia Intrínseca $\delta(o, e)$	Divergencia de Jeffreys $J(o, e)$	Discrepancia Intrínseca $\delta(o, e)$
Regr. Logística	0.048 bits	0.0235 bits	0.0582 bits	0.0280 bits
Red Neuronal	0.0271 bits	0.0134 bits	0.0314 bits	0.0153 bits

por lo que podemos afirmar que ambos modelos están muy bien calibrados. Pero el uso de estas medidas de información estadística nos permite además afirmar que las predicciones realizadas con el modelo de RL están a una distancia doble de las realizadas con el modelo de RN, por lo que la calibración del modelo de RN es

mejor. Esta afirmación es, de nuevo, imposible que se realice haciendo un análisis frecuentista.

2.3.- ANÁLISIS DE DECISIÓN:

Si se incorporan a un análisis formal de decisión desde la perspectiva del clínico que va a utilizar ambos modelos, tanto si se usan los índices de exactitud obtenidos mediante análisis frecuentista clásico como si se usan los obtenidos mediante nuestra propuesta bayesiana, la decisión dominante es siempre utilizar el modelo de RN para hacer nuestras predicciones de mortalidad en la UCIP con el índice PRISM. Sólo en las situaciones en las que la prevalencia de muerte sea superior a un umbral del 38,46% (si se usan los índices de exactitud frecuentistas) o del 40,7% (si se usan los bayesianos), unos valores muy alejados de las prevalencias de muerte de las cohortes de desarrollo y de validación, la decisión más adecuada es utilizar el modelo de RL para predecir la muerte.

Universitat d'Alacant

3.- CONSIDERACIONES FINALES

Por razones históricas, la estadística frecuentista ha sido la evasión prevalente de la inducción en ciencia. Era la universalmente enseñada, la única accesible en los paquetes de software de las calculadoras y las computadoras personales, y la que esperaban los editores de las revistas. Nick Metropolis y otros físicos desarrollaron la metodología de cadenas de Markov y Monte Carlo en 1953^{225,226}, su algoritmo fue generalizado en por Hastings en 1970²²⁷, y el muestreo de Gibbs se desarrolló en 1984²²⁸, pero estos métodos bayesianos no irrumpieron con fuerza en la ciencia estadística hasta los años 1990-1995. Y desde entonces las cosas están cambiando muy rápidamente, sobretodo tras la aparición en 1996 del software libre BUGS²²⁹, que utiliza cadenas de Markov y métodos de


Monte Carlo para generar muestras de la distribución a posteriori del modelo especificado. Ahora, que ya disponemos incluso de software libre y sencillo de utilizar (winBUGS 1.4.3 es la versión actual de BUGS para entorno Windows, que está disponible libre via Internet en la página web del proyecto WinBUGS²³⁰, y en español disponemos²³¹ del programa libre EpiDat 3.1, desarrollado el Servicio de Epidemiología de la Dirección Xeral de Saúde Pública da Consellería de Sanidade (Xunta de Galicia) en colaboración con la Unidad de Análisis de Salud y Sistemas de Información Sanitaria da Organización Panamericana de la Salud y de la Organización Mundial de la Salud [OPS-OMS]) está creciendo la popularidad de los métodos bayesianos en ciencia, que nos permiten hacer las cosas mucho mejor.

En esencia con nuestro trabajo no hemos pretendido ahondar en el debate sobre las dos grandes escuelas de la estadística contemporánea, sino sólo dar "una oportunidad para Bayes"²³², sobre todo en el terreno que le es más propicio: los métodos de diagnóstico médico. Demos "al César lo que es del César y a Dios lo que es de Dios"²³³.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante



***VIII. - Conclusiones y
proyección de futura
investigación.***



Universitat d'Alacant
Universidad de Alicante

PRIMERA: Nuestro trabajo confirma que, bajo el supuesto de intercambiabilidad y sin que se requiera de un muestreo aleatorio ni una justificación asintótica, es posible estimar la exactitud de un test diagnóstico mediante índices de rendimiento calculados con métodos estadísticos sencillos de naturaleza bayesiana.

SEGUNDA: Tanto el modelo de Regresión Logística como la Red Neuronal presentan una excelente capacidad de discriminación. Las estimaciones frecuentistas de los IC95% de los índices de exactitud presentan serios problemas de interpretación conceptual. Sin embargo, estimados con métodos bayesianos, los Intervalos del 95% de Credibilidad tienen una fácil y directa interpretación conceptual, sobretodo desde el punto de vista clínico. Así, sólo mediante métodos bayesianos hemos podido estimar que, si utilizamos en clínica el índice PRISM con el propósito de confirmar que se va a producir la muerte un niño en la UCIP, la probabilidad de que la Red Neuronal sea un buen método diagnóstico es altísima (0,87), muy superior a la de la Regresión Logística (0,04). Por el contrario, si lo utilizamos con el propósito de descartar que se vaya a producir la muerte, la probabilidad de que la Regresión Logística sea un buen método diagnóstico (0,92) es, superior a la de la Red Neuronal (0,66).

TERCERA: El análisis frecuentista no permite establecer la superioridad en la capacidad de discriminación de ninguno de los dos modelos: no se ha podido descartar la hipótesis de diferencia nula entre ambos (con una probabilidad de cometer error tipo I del 5%). Sin embargo, el análisis bayesiano nos permite obtener una información distinta y mucho más completa: la hipótesis más verosímil es la de que el modelo de Red Neuronal es más discriminante que el de Regresión Logística. De hecho la hipótesis de que el modelo de Red Neuronal sea mejor discriminante es tres veces más probable que la contraria.

CUARTA: Tanto el modelo de Regresión Logística como la Red Neuronal parecen estar muy bien calibrados. De nuevo el análisis frecuentista no nos ha permitido otra conclusión que la de asumir una buena calibración, porque en ambos casos no hemos podido descartar la hipótesis nula (con un valor alfa del 5%). Para realizar un análisis bayesiano de la calibración de métodos diagnósticos que estimen probabilidades (como es el caso de ambos modelos aquí estudiados) en el que no utilicemos pruebas de significación estadística, en este trabajo hemos propuesto emplear índices extraídos de la teoría matemática de la información. Con ellos puede estimarse "distancias" (en concreto distancias entrópicas), expresadas en bits de información, entre las distribuciones de probabilidad estimadas por cada modelo y los resultados finalmente observados. Con este análisis, tanto la Regresión Logística como la Red Neuronal se comportan como métodos de estimación probabilística muy bien calibrados.

QUINTA: Desde el punto de vista frecuentista no hemos podido estimar cual de los dos modelos presenta mejor calibración. Con el uso de las medidas de información estadística propuestas en este trabajo podemos afirmar que, respecto de los resultados de mortalidad finalmente observados, las predicciones realizadas con el modelo de Regresión Logística quedan a una distancia doble de las realizadas con el modelo de Red Neuronal, de donde podemos inferir que la calibración del modelo de Red Neuronal es mucho mejor. Esta afirmación es, de nuevo, imposible de realizar desde un punto de vista del análisis clásico frecuentista.

SEXTA: Si se incorporan a un análisis formal de decisión desde la perspectiva del clínico, que es el que va a utilizar ambos modelos, tanto si se usan los índices de exactitud obtenidos mediante análisis frecuentista clásico como si se usan los obtenidos mediante nuestra propuesta bayesiana, la decisión dominante es siempre utilizar el modelo de Red Neuronal para hacer nuestras predicciones de mortalidad en la UCIP con el índice PRISM. Sólo en contextos clínicos en los que la prevalencia de muerte en la UCIP se sitúe en un entorno del 40%, un valor muy alejado de la

prevalencia de muerte en las UCIs pediátricas españolas, la decisión más adecuada sería utilizar el modelo de Regresión Logística. Ello confirma de nuevo la superioridad del modelo de Red Neuronal, y debería hacernos replantear la práctica habitual de nuestras UCIPs.

A la luz de estas conclusiones, y desde nuestro particular punto de vista, sería necesario en el futuro seguir promoviendo en los estudios científicos el uso de técnicas de análisis capaces de aportar una información con muchos más matices y, sobre todo, con una interpretación mucho más directa, natural y racional de sus conclusiones que las técnicas basadas en las pruebas de significación estadística de las hipótesis. Las técnicas bayesianas permiten incorporar rigurosamente, durante el proceso de análisis, los puntos de vista o convicciones que se tenían antes del estudio: ese es el modo habitual como realizamos nuestros razonamientos en la vida real. Además, sus conclusiones se expresan en el lenguaje de la probabilidad, el lenguaje de la ciencia. Los cálculos necesarios son mucho más complicados, pero ya existen programas informáticos que los simplifican y los ponen al alcance de la mayoría de los investigadores preocupados por el rigor de sus métodos y la exactitud de sus resultados. A pesar de sus críticos y detractores, en este siglo que empezamos, el enfoque bayesiano debería consolidarse como el preponderante para posibilitar la consolidación de la racionalidad en la ciencia.



Universitat d'Alacant
Universidad de Alicante



XI. - Anexos.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

ANEXO I: TEOREMA DE RAMSEY-DE FINETTI

A.-Probabilidades, "Odds"²³⁴ y logit:

La ODDS es otra manera de expresar el concepto de probabilidad que se utiliza sobre todo en el mundo de las apuestas. Una Odds es el cociente entre una probabilidad y su complementaria: compara la verosimilitud de que un evento particular tenga lugar (que un determinado caballo gane una carrera o que una determinada enfermedad esté presente) [a], con la verosimilitud de que *no* tenga lugar [b], utilizando la expresión $o = a / b$. Para convertir una probabilidad en su odds, se utilizan las siguientes expresiones:

$$\text{Prob} = \text{Odds} / 1 + \text{Odds} \quad \text{y} \quad \text{Odds} = \text{Prob} / 1 - \text{Prob}$$

Así, la probabilidad 0.75 se convierte en la Odds 0.75/0.25. El logaritmo neperiano (natural) de una Odds se denomina "LOGIT", y así

$$\text{Logit}(p) = \ln [p/1-p]$$

Las probabilidades son números del intervalo (convexo) [0,1], pero las odds pertenecen al intervalo $[0, \infty[$. Para obtener expresiones numéricas manejables cuando utilizamos las Odds, podemos multiplicar o dividir ambos elementos del cociente a/b por el mismo número, sin cambiar su significado. Así, una probabilidad 0.75 se corresponde con una Odds 0.75 / 0.25 que, si dividimos ambos miembros por 0.25, se convierte en 3/1.

Para convertir fácilmente la expresión de una odds a/b en una probabilidad, conviene dividir el numerador [a] por la suma de numerador y denominador [a+b]. Así, la odds 4/1 se convierte en la probabilidad $4/4+1 = 4/5$, es decir 0.8 (80%).

Si utilizamos el concepto bayesiano (subjetivo) de probabilidad como la medida de la incertidumbre o grado de creencia (racional), también podemos usar las odds correspondientes en términos del lenguaje de las apuestas. Así $P_i(E/H)$ puede interpretarse como el máximo precio que el individuo *i* está dispuesto a pagar por participar en una apuesta, jugada en las condiciones H, en la que se obtiene un premio unidad si, y solamente si, tiene lugar el evento E. Una persona coherente que afirme que la moneda lanzada tiene probabilidad 0.5 de obtener cara, debería estar dispuesta a pagar 50 euros por participar en una apuesta que le diera derecho a cobrar 100 euros si al lanzar la moneda se obtiene cara. Recíprocamente, sea una apuesta que da un premio de 100 euros si nuestro equipo de fútbol gana la liga y la Copa de la UEFA; si el máximo precio que una persona está dispuesta a pagar por participar en esa apuesta es 70 euros, la probabilidad que esta persona otorga a que su equipo gane ambas competiciones es 0.7 (70%)⁴⁶.

B.-Coherencia de una apuesta:

Cualquier corredor de apuestas sabe que nunca se debe jugar en apuestas tramposas, apuestas contra él mismo. Los ingleses utilizan la expresión "apuesta

del Holandés" (Dutch Book) para designar tales apuestas: aquellas en las que, sea cual sea el precio que pague el jugador por participar, siempre hay una pérdida neta (ganancia neta de la banca). Se llama **coherente** a una apuesta que no dé ninguna posibilidad a que se produzca Dutch Book²³⁵. Sólo hemos de participar en apuestas coherentes, para evitar la posibilidad de que se nos haga trampa y "nos tomen el pelo".

Sea una apuesta sobre la ocurrencia de un acontecimiento E en la que intervienen el "señor A" y "la banca". El "señor A", paga un precio $[p(E)*S]$ por participar y la banca pagará a este un premio $[S]$ si se produce el evento E. La medida que caracteriza a una apuesta se llama "**cociente de apuesta**" $p(E)$, y es el resultado de dividir el precio entre el premio [precio/premio]: el porcentaje del premio que se ha de pagar por apostar. La ganancia del "señor A" se define por: $GA = \text{premio} - \text{precio} = S - [p(E)*S]$, y la de "la banca" por: $GB = \text{precio} - \text{premio} = [p(E)*S] - S$. El "señor A" está claramente apostando a favor del evento E con una Odds = $p(E)*S / [S - p(E)*S] = p(E) / [1 - p(E)]$; y "la banca" está apostando en contra del evento E con la Odds recíproca = $[1-p(E)] / p(E)$. La apuesta es coherente sólo si el precio que paga el "señor A" y el premio que paga "la banca" se eligen de tal manera que es imposible que alguno de los dos gane siempre ($G = 0$). Si "la banca" elige los premios de tal manera que siempre gane, sea cual sea el precio que pague el "señor A", la apuesta es una incoherente "apuesta del holandés". Ramsey⁵³ y De Finetti⁵⁶, independientemente, demostraron un hecho sorprendente: es condición necesaria y suficiente para que la apuesta (o serie de apuestas) en que juguemos sea **coherente**, que nuestras apuestas se expresen en medidas $p(E)$ que cumplan las reglas que definen a las probabilidades.

C.-Los axiomas de la probabilidad (versión bayesiana):

La probabilidad es siempre condicionada, y depende de dos elementos: el evento sobre el cual sentimos incertidumbre y las condiciones en las que estamos realizando la medida de esa incertidumbre. Escribimos $p(E/H)$ para expresar nuestra probabilidad de que se produzca el evento E cuando sabemos, o asumimos como cierto, que se cumple la condición H. Es nuestra probabilidad de E dado H.

a) Regla 1 (**convexidad**): para cualquier E y H,

$$0 \leq p(E/H) \leq 1, \text{ y } p(E/E) = 1$$

b) Regla 2 (**ley aditiva** de eventos excluyentes): si E_1 y E_2 son eventos mutuamente excluyentes (no pueden darse a la vez) dado H,

$$p(E_1 \cup E_2 / H) = p(E_1 / H) + p(E_2 / H)$$

En concreto, si E_1, \dots, E_n son varios eventos mutuamente excluyentes y exhaustivos (al menos alguno de ellos ha de acontecer) dado H,

$$p(E_1 / H) + \dots + p(E_n / H) = 1$$

c) Regla 3 (**ley multiplicativa**): para cualquier A, E y H

$$p(A \& E / H) = p(A / E \& H) * p(E / H), \text{ siendo}$$

(A&E) el evento que ocurre si, y sólo sí, ocurren conjuntamente A y E.

En concreto, si A y E son sucesos independientes, dado H [$p(A/E \& H) = p(A/H)$], entonces $p(A \& E / H) = p(A / H) * p(E / H)$

La ley multiplicativa puede ser expresada de manera equivalente con la definición de probabilidad condicional:

$$p(A / E \& H) = p(A \& E / H) / p(E / H), \text{ siempre que } p(E / H) > 0$$

Como resultado inmediato importante de las tres reglas, se obtiene la llamada *Ley de Marginalización* (también conocida con el curioso nombre de ley para "Extender la Conversación" o "Extender el Argumento"). Se utiliza mucho para describir una distribución conjunta global sobre todas las posibles combinaciones de eventos. Supongamos los eventos $(E \& B)$ y $(E \& \text{no}B)$. Como $(E \& B)$ y $(E \& \text{no}B)$ son mutuamente excluyentes y forman conjuntamente el evento E , tenemos [Ley aditiva de eventos excluyentes]:

$$p(E) = p(E \& B) + p(E \& \text{no}B)$$

y usando la ley multiplicativa

$$p(E) = p(E / B) p(B) + p(E / \text{no}B) p(\text{no}B)$$

D.-Prueba del Teorema de Ramsey-De Finetti^{52,63,236}:

a) Concepto de "cociente condicional de una apuesta": La prueba del teorema necesita que se cumpla la definición de cociente condicional de una apuesta. El cociente condicional de una apuesta $p(E/H)$ del "señor A" contra "la banca" es el cociente de una apuesta [precio/premio] que se realiza entre ellos sólo bajo la condición de que, si H no ocurre, la apuesta termina y "la banca" devuelve todo el dinero pagado por el "señor A" como precio por apostar.

b) Prueba para la Regla 1 (**convexidad**):

❖ Coherencia \rightarrow Regla 1:

Si el "señor A" elige para el suceso seguro una $p(E/E) > 1$, y paga por él un precio [$p(E/E)*S$], "la banca" siempre gana ($GB = [p(E/E)*S] - S > 0$) poniendo cualquier premio $S > 0$. Si el "señor A" elige una $p(E/E) < 1$, "la banca" siempre gana poniendo cualquier premio $S < 0$ (negativo). Para que la apuesta sea coherente, el "señor A" deberá apostar a favor de un evento seguro con una $p(E/E) = 1$. Sea ahora un evento arbitrario no seguro E/H . Si el "señor A" elige un $p(E/H) > 1$, "la banca" siempre gana poniendo cualquier premio $S > 0$. Si el "señor A" elige $p(E/H) < 0$ (negativo), "la banca" puede ganar si el premio es < 0 (negativo). Así que, exigiendo coherencia, el "señor A" debe asignar probabilidades a favor del evento $0 \leq p(E/H) \leq 1$

❖ Regla 1 \rightarrow Coherencia:

Si el "señor A" elige para el suceso seguro una $p(E/E) = 1$ no hay manera de que "la banca" gane dinero ($GB = [1*S] - S = 0$). El premio, sea cual sea el signo, pasa simplemente de una mano a otra y de nuevo de vuelta al

primero. Para un acontecimiento arbitrario no seguro E/H , "la banca" no puede elegir un valor de premio S que conlleve ganancia segura si el "señor A" paga por apostar un precio $[p(E/H)*S]$ con $0 \leq p(E/H) \leq 1$.

c) Prueba para la Regla 2 (**ley aditiva** de eventos excluyentes):

❖ **Coherencia** → Regla 2:

Supongamos que en una serie de n apuestas, el "señor A" paga una serie de precios $[p(E_1/H)*S_1], [p(E_2/H)*S_2], \dots, [p(E_n/H)*S_n]$ por participar y "la banca" paga una serie de premios S_1, S_2, \dots, S_n en cada una de ellas si ocurren una serie de eventos E_1, E_2, \dots, E_n mutuamente excluyentes y exhaustivos. Sea que "la banca" paga el mismo premio en cada una de las apuestas: $S_1 = S_2 = \dots = S_n = S$.

Si ocurre el evento E_i , la ganancia neta de "la banca" GB viene dada por:

$$GB = p(E_1/H)*S_1 + p(E_2/H)*S_2 + \dots + p(E_n/H)*S_n - S_i$$

y como $S_1 = S_2 = \dots = S_n = S$, entonces

$$GB = S * [p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) - 1]$$

Si, para el "señor A" $p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) > 1$, "la banca" gana siempre poniendo cualquier premio $S > 0$. Si para el "señor A" $p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) < 1$, la banca gana siempre poniendo cualquier premio $S < 0$ (negativo). Para que la serie de apuestas sea coherente ($GB = 0$), el "señor A" debe elegir $p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) = 1$.

❖ Regla 2 → **Coherencia**:

Hemos visto que $GB = p(E_1/H)*S_1 + p(E_2/H)*S_2 + \dots + p(E_n/H)*S_n - S_i$, y como $S_1 = S_2 = \dots = S_n = S$, entonces

$$GB = S * [p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) - 1]$$

Desde la Regla 2, tenemos que $p(E_1/H) + p(E_2/H) + \dots + p(E_n/H) = 1$ para el "señor A". Ello hace que la serie de apuestas sea coherente ($GB = 0$).

d) Prueba para la Regla 3 (**ley multiplicativa**):

❖ **Coherencia** → Regla 3:

Sea una apuesta sobre la ocurrencia de tres eventos cualesquiera A, E y H . La apuesta se rige por los siguientes cuyos cocientes de apuesta: $p(A\&E/H)$, $p(A/E\&H)$ y $p(E/H)$. Ello significa que:

- El "señor A" paga un precio $[p(A\&E/H)*S]$, y recibe de "la banca" un premio S , si ocurre el evento conjunto $A\&E$ siempre que ocurra primero H
- El "señor A" paga un precio $[p(A/E\&H)*S']$, y recibe de "la banca" un premio S' si aparece el evento A siempre que ocurra primero el evento conjunto $E\&H$, y
- El "señor A" paga un precio $[p(E/H)*S'']$, y recibe de "la banca" un premio S'' , si ocurre el evento E siempre que ocurra primero el H .

Una vez ha ocurrido H , pueden ocurrir tres cosas:

- 1) Ocurren conjuntamente A y E ($A\&E$), y la ganancia de "la banca" es :

$$GB1 = [p(A\&E/H) - 1]*S + [p(A/E\&H) - 1]*S' + [p(E/H) - 1]*S''$$

2) No ocurre A, pero ocurre E, y la ganancia de "la banca" es :

$$GB2 = p(A\&E/H) * S + p(A/E\&H) * S' + [p(E/H) - 1] * S''$$

3) No ocurre E, y la ganancia de "la banca" es :

$$GB3 = p(A\&E/H) * S + p(E/H) * S''$$

Supongamos que "la banca" elige $S = +1$, $S' = -1$ y $S'' = -p(A/E\&H)$, luego

$$GB1 = [p(A\&E/H) - 1] + [1 - p(A/E\&H)] + p(A/E\&H) - [p(A/E\&H) * p(E/H)] \\ = p(A\&E/H) - p(A/E\&H) * p(E/H)$$

$$GB2 = p(A\&E/H) - p(A/E\&H) - p(A/E\&H) * p(E/H) + p(A/E\&H) = \\ = p(A\&E/H) - p(A/E\&H) * p(E/H)$$

$$GB3 = p(A\&E/H) - p(A/E\&H) * p(E/H)$$

es decir que $GB1 = GB2 = GB3$ y son positivas (> 0) salvo que $p(A\&E/H) \leq p(A/E\&H) * p(E/H)$. Igualmente, si "la banca" elige $S = -1$, $S' = +1$ y $S'' = +p(A/E\&H)$, $GB1 = GB2 = GB3$ y son positivas (> 0) a no ser que $p(A\&E/H) \geq p(A/E\&H) * p(E/H)$. Para que se dé la coherencia [$GB1 = GB2 = GB3 = 0$], el "señor A" debe elegir $p(A\&E/H) = p(A/E\&H) * p(E/H)$

❖ Regla 3 → Coherencia:

Siguiendo el mismo razonamiento, si el "señor A" elige $p(A\&E/H) = p(A/E\&H) * p(E/H)$, la ganancia de "la banca" es $GB1 = GB2 = GB3 = p(A\&E/H) - p(A/E\&H) * p(E/H) = 0$. Si se sigue la Regla 3, se obtiene coherencia.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

ANEXO II: INTERCAMBIABILIDAD Y TEOREMA DE REPRESENTACIÓN

El estudio de las probabilidades intercambiables es un campo tremendamente vivo y productivo de la investigación matemática de nuestros días. Pero, en esencia, la noción de *intercambiabilidad* corresponde a una idea muy simple que se le ha ocurrido a muchos: la de considerar que *secuencias* de eventos binarios de Bernoulli, cada una de las cuales tiene el mismo tamaño n y el mismo número de éxitos r , son secuencias a las que se puede asignar la misma probabilidad. El caso especial más famoso es la *ley de sucesiones* de Laplace: sea p la probabilidad desconocida de éxito de un evento simple. De acuerdo con Laplace, a priori uno no tiene ninguna razón para pensar que un valor de p sea más probable que otro. Esta ignorancia sobre el valor de p se representa por una distribución uniforme sobre todos los valores de p . Las probabilidades posteriores pueden calcularse con el teorema de Bayes: la probabilidad de un nuevo éxito, una vez se ha realizado un experimento de r éxitos en n intentos, es $(r+1)/(n+2)$. Es un ejemplo de asignación de probabilidades intercambiables, pero hasta de Finetti nadie había caído en que ello vá más allá de los experimentos con secuencias de eventos binarios independientes.

Consideremos un fenómeno aleatorio como aquel en el que se puede hacer -o al menos imaginar hacer- un experimento: generar una secuencia de un número n de intentos repetidos y en el que se considera que *el orden* en el que aparecen los valores que toma la variable aleatoria (p. ej. el orden de éxitos y fracasos en una sucesión de n pruebas de Bernoulli) puede achacarse estrictamente al azar. La condición matemática de intercambiabilidad se establece sobre secuencias o sucesiones de valores de la variable aleatoria (eventos), y consiste en que la probabilidad de obtener esa secuencia es la misma independientemente del orden en que se obtengan esos valores o eventos.

Sea un fenómeno aleatorio de Bernoulli. Siguiendo a de Finetti⁹⁴, llamaremos $w_r^{(n)}$ a la probabilidad de obtener r éxitos en una secuencia de n intentos intercambiables. Si los sucesos son *independientes* con probabilidad p , la probabilidad de obtener r éxitos en cualquier secuencia de tamaño n es $p^r(1-p)^{n-r}$. Como el número de las posibles tales secuencias es $\binom{n}{r}$, la probabilidad de que la frecuencia de éxitos sea r/n es $\binom{n}{r} p^r(1-p)^{n-r}$. Como puede apreciarse, esta expresión es *independiente del orden* de la secuencia, así que independencia mas eventos equiprobables es un caso particular de intercambiabilidad.

Supongamos ahora que tenemos m urnas con bolas blancas y negras, siendo p_i la probabilidad de obtener una bola blanca en la urna i . Asumamos que la probabilidad a priori de elegir la urna i es a_i . En n repeticiones del experimento, primero elegiremos la urna de acuerdo con las probabilidades a_i y luego

extraeremos una bola, con reemplazamiento. Con la fórmula de la probabilidad total obtenemos que la probabilidad de sacar r bolas negras en n intentos es:

$$w_r^{(n)} = \binom{n}{r} \sum a_i p_i^r (1-p_i)^{n-r}.$$

Esta mezcla de eventos de Bernoulli nos da una probabilidad que es *intercambiable*, esto es, independiente del orden de la secuencia. Sea ahora un mecanismo aleatorio de extraer un número $p \in [0,1]$ de acuerdo con una densidad de probabilidad $f(p)$. Entonces la probabilidad conjunta es:

$$w_r^{(n)} = \binom{n}{r} \int_0^1 p^r (1-p)^{n-r} f(p) dp.$$

Esta probabilidad es, de nuevo, *intercambiable*.

Así, leído de derecha a izquierda este teorema parece un resultado matemático puro más de la teoría de probabilidades: si y_1, y_2, \dots, y_n es una secuencia de variables aleatorias probabilísticamente independientes entre sí (dadas las condiciones k) e idénticamente distribuidas (I.I.D.) -es una muestra aleatoria-, cada una con verosimilitud $p(y_i / \psi)$, su distribución conjunta (condicionada en ψ) es [Ley multiplicativa]:

$$p(y_1, y_2, \dots, y_n / \psi) = \prod_{i=1}^n p(y_i / \psi).$$

Por ello, su distribución marginal dada una distribución $p(\psi)$ es [Marginalización]:

$$p(y_1, y_2, \dots, y_n) = \int_0^1 \prod_{i=1}^n p(y_i / \psi) p(\psi) d\psi$$

y ello implica que son intercambiables entre sí: el orden en el que se han extraído no importa para obtener su distribución conjunta. Cualquier muestra aleatoria genera una secuencia que es intercambiable, pues $\prod_{i=1}^n p(y_i / \psi)$ es obviamente invariante bajo permutaciones.

Pero el teorema de representación de de Finetti demuestra que el inverso es cierto: que para cualquier ley probabilística intercambiable para todo n , **existe una distribución única $F(p)$** tal que:

$$w_r^{(n)} = \binom{n}{r} \int_0^1 p^r (1-p)^{n-r} dF(p),$$

y si $F(p)$ tiene densidad $f(p)$, entonces

$$w_r^{(n)} = \binom{n}{r} \int_0^1 p^r (1-p)^{n-r} dF(p) = \binom{n}{r} \int_0^1 p^r (1-p)^{n-r} f(p) dp.$$

Así que el éxito notable del teorema es su argumentación de izquierda a derecha: nos muestra como eliminar las probabilidades objetivas desconocidas p en favor de las probabilidades subjetivas. Una sucesión o secuencia de variables aleatorias sobre las que establezcamos un **juicio subjetivo** de que son intercambiables entre sí (en las condiciones K) -esto es, de que la probabilidad de obtener esa secuencia es independiente del orden-, pueden ser consideradas como variables I.I.D.; esto es, como una muestra aleatoria generada "al azar" -cuyo orden depende sólo del azar- por *algún* modelo probabilístico caracterizado por un

parámetro ψ . El teorema resulta de la aplicación del método de funciones características a las probabilidades intercambiables, ya que de hecho es un subproducto del resultado matemático central de las funciones características: la determinación de la distribución, en el límite cuando $n \rightarrow \infty$, de la frecuencia relativa de una secuencia intercambiable de eventos. Y es que este parámetro ψ resulta ser el límite (cuando $n \rightarrow \infty$) de alguna función de los datos $f(y_1, y_2, \dots, y_n)$, y sobre este parámetro, a su vez, **debe necesariamente** establecerse alguna función de probabilidad "a priori".

La idea de ensayos repetidos relaciona este problema con la probabilidad frecuentista, pero el objetivo de de Finetti es mostrarnos como, y en que sentido, es posible realizar una inferencia desde datos de frecuencias relativas a probabilidades entendidas según su aproximación subjetiva. Primero debe hacerse explícito un juicio (subjetivo) sobre la intercambiabilidad de los eventos aleatorios, y segundo es necesario establecer una distribución a priori particular sobre el parámetro que caracteriza el modelo probabilística que está generando los datos. De otro modo, tal inferencia es imposible. Desde un juicio subjetivo sobre cantidades observables se deriva todo el armamento de la teoría de la probabilidad válido para variables I.I.D. -independencia condicional-, la existencia de distribuciones "a priori" sobre los parámetros, y la existencia (verificable sólo asintóticamente) de esos parámetros definidos como cantidades fijas e invariables.

La prueba de este teorema queda fuera de los límites de este estudio, pero está disponible en el libro de Bernardo y Smith⁸⁸. La demostración para el modelo binomial de variables aleatorias dicotómicas de Bernoulli se debe a De Finetti^{94,237}, que estableció la distribución límite de una frecuencia relativa cuando se asume intercambiabilidad. Aunque la demostración moderna más rigurosa puede consultarse en Heath y Sudderth²³⁸, una prueba muy simple se encuentra en Lindley y Phillips⁹⁷. La demostración de Bernardo para el modelo general utiliza instrumentos de la teoría general de la medida, aunque una prueba parcial se encuentra en Chow y Teicher²³⁹.

ANEXO III: TEST BAYESIANO DE HIPÓTESIS: RAZON DE VEROSIMILITUDES y FACTOR DE BAYES

La cantidad usada para testar hipótesis en estadística bayesiana se denomina **Factor de Bayes**. Supóngase que queremos testar H_0 frente a H_1 . Como bayesianos, debemos primero especificar las probabilidades a priori para H_0 y H_1 , que llamaremos $p(H_0)$ y $p(H_1)$. Denotaremos por D a los datos recogidos mediante un experimento, y llamaremos $p(D/H_0)$ y $p(D/H_1)$ a las probabilidades posteriores de H_0 y H_1 , respectivamente. Pues bien, el Factor de Bayes (FB) *a favor de H_0 (y en contra de H_1)* se define como el cociente entre dos razones: la razón de probabilidades posterior/prior de H_0 , dividida por la de H_1 . Así:

$$FB = \frac{p(H_0/D)/p(H_0)}{p(H_1/D)/p(H_1)} = \frac{p(H_0/D)/p(H_1/D)}{p(H_0)/p(H_1)} \quad [1]$$

En el caso particular de que H_0 y H_1 sean dos hipótesis complementarias, es decir $p(H_1) = 1-p(H_0)$ y $p(D/H_1) = 1-p(D/H_0)$, esta definición del BF coincide con la **Odds Ratio** de Odds a posteriori a favor de H_0 partido por la Odds a priori a favor de H_0 .

Sabemos además, por el Teorema de Bayes, que:

$$p(H_0/D) = \frac{p(D/H_0)p(H_0)}{p(D/H_0)p(H_0) + p(D/noH_0)p(noH_0)} = \frac{p(D/H_0)p(H_0)}{p(D)}$$

Y una fórmula similar para $p(H_1/D)$. Substituyendo esto en [1], tenemos que el Factor de Bayes *a favor de H_0 (y en contra de H_1)* es:

$$FB = \frac{p(D/H_0)}{p(D/H_1)}$$

El Factor de Bayes, así definido, tiene una interpretación muy elegante. **Cuanto mayor es el FB, mayor es la evidencia de los datos a favor de H_0 (y en contra de H_1).**

De una manera metamáticamente más formal, podemos expresar el factor de Bayes en términos de densidades de probabilidad. Supóngase que nuestro espacio de parámetros es Θ y queremos testar

$$H_0: \theta \in \Theta_{H_0} \text{ frente a } H_1: \theta \in \Theta_{H_1} ,$$

donde $\Theta = \Theta_{H_0} \cup \Theta_{H_1}$ y $\Theta_{H_0} \cap \Theta_{H_1} = \emptyset$.

Entonces, el Factor de Bayes *a favor de H_0 (y en contra de H_1)* se define como

$$FB = \frac{p(D/H_0)}{p(D/H_1)} = \frac{\int_{\Theta_{H_0}} p(D/\theta, H_0)\pi(\theta/H_0)d\theta}{\int_{\Theta_{H_1}} p(D/\theta, H_1)\pi(\theta/H_1)d\theta}$$

Donde Θ_{H_0} es el espacio de parámetros bajo H_0 , y Θ_{H_1} es el espacio de parámetros bajo H_1 .

Consecuencias inmediatas importantes de esta definición del Factor de Bayes son:

1. Las verosimilitudes de los datos bajo las hipótesis $p(D/H_0)$ y $p(D/H_1)$ se obtienen por INTEGRACIÓN sobre el espacio de parámetros, **NO MAXIMIZANDO**.
2. En el caso de que estemos testando dos hipótesis complementarias, el FB coincide con la Odds Ratio posterior versus prior a favor de H_0 .
3. En el caso particular de queramos testar hipótesis simples, el Factor de Bayes se reduce al cociente de verosimilitudes. Así, para

$$H_0 : \theta = \theta_0 \quad \text{frente a} \quad H_1 : \theta = \theta_1$$

$$FB = \frac{p(x/\theta = \theta_0)}{p(x/\theta = \theta_1)}$$

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

ANEXO IV: TEORIA DE LA INFORMACIÓN

La teoría matemática de la Información fué iniciada por Nyquist^{240,241} (1924) y Hartley²⁴² (1928), y desarrollada desde los trabajos de A. N. Kolmogorov y Norbert Wiener²⁴³ (1948) sobre todo por Claude Elwood Shannon (1948)⁵⁸, ingeniero y matemático que nació en 1916 en Michigan (EE.UU.), trabajó en el Massachusetts Institute of Technology (el famoso MIT) y murió en 2001. Primitivamente denominada "teoría matemática de la comunicación" Shannon la publicaría inicialmente, mientras trabajaba como ingeniero de telecomunicaciones en el Bell Telephone Laboratorios, como una prolongación de sus trabajos sobre criptología para el ejército de EE.UU. durante la segunda guerra mundial. Versaba sobre la transmisión de la información y las mejoras de la codificación, sin tener en cuenta el significado de los mensajes transmitidos. Intentando ligar la forma en que se codifica un mensaje a la dificultad que debe vencer quien lo intercepta y pretende descifrarlo, Shannon proponía una forma unívoca de cuantificar la información que puede contener un mensaje. Con esta idea genial pronto sobrepasó el campo de la ingeniería de telecomunicaciones hasta alcanzar el enorme desarrollo que posee en nuestros días. Actualmente se ha aplicado a todas las formas posibles de intercambio de información, pues la información ha perdido su carácter subjetivo y se ha tornado mensurable, cualquiera que sea su origen: se ha podido cuantificar la información que contiene una lengua, una melodía, una secuencia de ADN, etc... Con independencia del soporte, en nuestros días la información ha alcanzado tal estatus primordial, que nuestra era ha pasado a denominarse por muchos la era de la información²⁴⁴.

A. -Probabilidad e Información:

Sea un acontecimiento cualquiera del mundo real -por ejemplo la expulsión de Zidane de la final del mundial de Alemania 2006, o la muerte de un niño ingresado en la UCIP-, y un observador imparcial²⁴⁵. La ocurrencia de tal acontecimiento divide la historia de la realidad en dos partes, la de **antes** y la de **después**. Para el observador curioso, ambas etapas tienen un interés muy diferente: verosimilitud y probabilidad tienen que ver con el **antes**, sorpresa e información con el **después**.

Lo más característico del **antes** es que el acontecimiento todavía no ha ocurrido, por lo que el observador se encuentra en un ambiente de incertidumbre en el que sólo le queda limitarse a "apostar": jugar a imaginar y a calcular expectativas. Durante el **antes**, el acontecimiento imaginado sólo puede ser verosímil, capaz de hacerse realidad. Pero la verosimilitud tiene límites: un acontecimiento que no ha ocurrido nunca puede ser más verosímil que algo cierto, ni menos verosímil que algo imposible. Lo cierto ha accedido necesariamente a la realidad y lo imposible nunca podrá acceder. Ello sugiere algo muy interesante para una magnitud científica: la verosimilitud tiene grados. La ocurrencia de un evento es muy verosímil si se acerca a la certeza (mañana será otro día) y es poco si se acerca a la imposibilidad (la gata parirá una gaviota). Pero en la realidad de nuestro

mundo es fácil convencerse de que, de hecho, tanto la certeza como la imposibilidad son dos situaciones ideales, no reales. Es tan difícil imaginar algo totalmente cierto como algo totalmente imposible. Así, por muy seguro que parezca el acontecimiento "mañana será otro día", hay que admitir que, para todos aquellos planetas que ya han desaparecido hubo un día (sólo uno, de acuerdo) en el cual esta afirmación fue falsa. No tenemos, por tanto, la certeza absoluta de que mañana saldrá el sol. Análogamente, "la gata parirá una gaviota" suena efectivamente del todo imposible, pero si no fuera por la realización de imposibilidades como esta la evolución biológica no se hubiera producido. Los extremos de la verosimilitud son pues sólo relevantes en tanto que situaciones límite en enmarcan a lo que en realidad nos interesa: eventos reales con sus valores intermedios. En un ambiente de incertidumbre, la verosimilitud es, pues, una magnitud mensurable i merece por ello rango científico. Se llama *probabilidad* de un acontecimiento -¡por el mero hecho de no haber ocurrido!- a la medida (número real entre 0 y 1) de esa verosimilitud, a la medida de su distancia a la certeza (valor 1) o a la imposibilidad (valor 0). Cuando los eventos ya han ocurrido son siempre ciertos (no hay incertidumbre) y su verosimilitud es máxima. No tiene sentido hablar de probabilidad de un evento *después* de su ocurrencia: el concepto de probabilidad está asociado a la incertidumbre previa a la ocurrencia.

Lo más característico del *después* es que el acontecimiento ya ha ocurrido. Cuando un evento ocurre, la incertidumbre y las expectativas se evaporan, y al observador sólo le queda una cosa por hacer: sorprenderse. Sorprenderse mucho (*después*) si el acontecimiento era (*antes* de ocurrir) muy poco probable, o al revés. Estos son los límites de la sorpresa: en el límite inferior, casi nadie se sorprende porque mañana comience otro día (sorpresa próxima a cero); en el límite superior, ¿quién podría disimular su sorpresa si asistiera al parto de una gaviota por una gata? (sorpresa casi infinita). En realidad la sorpresa constituye el cambio que se produce en nuestro estado de ánimo por la ganancia de información que acontece cuando se produce el evento: un cambio en nuestro conocimiento provocado por la información derivada del evento. La *información* como expresión de nuestra sorpresa. La medida de esa información, *después* de ocurrir el evento, depende de la probabilidad que este acontecimiento tenía asignada por nosotros en situación de incertidumbre, *antes* de su ocurrencia. Es decir, cualquier acontecimiento que ha ocurrido tiene asignado -¡por el mero hecho de haber acontecido ya!- un número que mide (en *bits*) la información que nos ha proporcionado su ocurrencia. Esta medida varía entre el cero (la información suministrada por un acontecimiento cierto) y el infinito (acontecimiento imposible). La información queda así asociada a un acontecimiento y a una mente capaz de valorar y constatar su ocurrencia, una vez que ésta se ha producido.

B. - Medida de la Información:

Tenemos ya una intuición clara del concepto de información proporcionada por la ocurrencia de un evento. Pero necesitamos una definición matemáticamente

rigurosa -completa y coherente- de una magnitud que satisfaga todos esos requisitos intuitivos y, sobretodo, que sea capaz de medirla en unas unidades pre-establecidas. Shannon la encontró, con un razonamiento que vamos a resumir a continuación.

Sea un acontecimiento A . Antes de su ocurrencia, dicho acontecimiento tiene asignada una probabilidad $P(A) = p$. Después de su ocurrencia, por el simple hecho de que esta se ha producido, el acontecimiento A nos aporta una información. Se busca una función $I(A)$ que satisfaga nuestra intuición y toda una serie de propiedades o requisitos que esta función debe cumplir, que son:

B.1.- La información es función de la probabilidad:

En primer lugar el acontecimiento A no es un ente matemático, pero la probabilidad p que tenía antes de ocurrir sí que es una entidad matemática bien definida por el concepto de Kolmogorov¹⁸, es decir $0 \leq p \leq 1$. Así que lo que buscamos es una función del concepto matemático de probabilidad: una función $I(p)$ definida como la información proporcionada por la ocurrencia de un acontecimiento A , la probabilidad del cual, antes de ocurrir, era p .

B.2.- $I(p)$ es una función *decreciente* de la probabilidad:

Está clarísimo en nuestro concepto intuitivo que cuanto menos probable es el acontecimiento, más información nos aporta una vez se produzca finalmente su ocurrencia. La información de un acontecimiento imposible es infinita, es decir $I(0) = \infty$, y la de un evento cierto es nula, esto es $I(1) = 0$.

B.3.- $I(p)$ es una función *continua* decreciente de la probabilidad:

La continuidad es una propiedad muy apreciada para caracterizar a las magnitudes científicas. Se trata no sólo de que la función $I(p)$ crezca con la disminución de p , sino que a un cambio minúsculo de la probabilidad del evento A le corresponda un efecto igualmente minúsculo en la información que se gana con la ocurrencia de A . Esta es una propiedad importante, ya que el concepto intuitivo de cambio se representa matemáticamente con el concepto de derivada, y la continuidad es, justamente, una condición necesaria para que una función sea diferenciable. Por ello exigimos continuidad.

B.3.- $I(p)$ es una función *aditiva* continua decreciente de la probabilidad:

La aditividad es otra propiedad de gran prestigio para las medidas de rango científico. Hasta ahora nos hemos referido a la información asociada a un evento aislado, pero es fácil apreciar el mérito que representa poder evaluar ágilmente la información asociada a un conjunto de acontecimientos. En efecto sean dos eventos independientes A y B de probabilidades $P(A) = p$ y $P(B) = p'$. La ocurrencia simultánea de ambos acontecimientos es otro evento que, si son independientes, tiene asociada una probabilidad previa a su ocurrencia

$$P(A \cap B) = P(A) \cdot P(B) = p \cdot p'$$

La ocurrencia de este evento compuesto también aportará, cuando se produzca, una información $I(A \cap B)$. Exigir aditividad para la información implica exigir que la información obtenida tras la ocurrencia simultánea de dos acontecimientos independientes sea la suma de las informaciones aportadas por cada uno de los eventos por separado

$$I(A \cap B) = I(p \cdot p') = I(p) + I(p')$$

El **primer teorema** de la teoría matemática de la información demuestra que existe una, y sólo una, función de la probabilidad que sea aditiva continua y decreciente: la logarítmica. Así:

Si $I: [0,1] \rightarrow \mathbb{R}^+$ satisface dos condiciones

- a) $I(p)$ es una función continua decreciente de p
- b) $I(p \cdot p') = I(p) + I(p')$ para todo $p, p' \in [0,1]$

entonces

$$I(p) = \mu \log_2(1/p) = -\mu \log_2 p$$

donde μ es una cierta constante positiva.

La demostración excede el ámbito de este estudio, pero está accesible en la literatura sobre el tema^{246,245}. Es bella, ingeniosa y sencilla, y proporciona además el significado del parámetro μ . Ya que el logaritmo es sólo un exponente (un número sin dimensiones), esta definición nos permite definir las unidades en que vamos a medir la magnitud información, que asignaremos justamente a μ :

$$I(p) = -\mu \log_2 p = \mu \log_2(1/p) \leftrightarrow \mu = I(1/2)$$

Es decir que la unidad de información, es la información aportada -después de su ocurrencia- por un acontecimiento la probabilidad del cual, antes de ocurrir, era $p = \frac{1}{2}$. A esta unidad de información se le llama **bit** (contracción de *binary unit*), en honor a la arbitraria base 2 del logaritmo que se usa en la definición. Una vez demostrado el teorema, la magnitud información queda bien definida:

$$I(p) = \log_2(1/p) \text{ bits} = -\log_2 p \text{ bits}$$

Sabiendo que una de las principales propiedades de los logaritmos es

$$\log_a Y = \log_b Y / \log_b a$$

podemos medir la información expresándola con logaritmo neperiano (base e), y entonces la unidad de información se llama nit (*natural unit*).

C. - Sistema fuente de la información: Entropía de Shannon

En general, al estudiar la naturaleza, no encontramos acontecimientos aislados, sino conjuntos de estos acontecimientos. De todo lo tratado en el apartado anterior surge, como generalización natural, la idea de medir la información que, globalmente, es capaz de producir un sistema generador de eventos, una fuente de información. Lanzar un dado es un sistema generador de seis posibles eventos independientes, hacer girar una ruleta, de treinta y seis. Pero en ese mismo sentido, un idioma se puede considerar como un generador de letras, palabras o fonemas; la música, un generador de sonidos o acordes; la pintura, uno de colores; y un ecosistema, uno de especies, géneros e individuos. Del concepto de

información proporcionada por la ocurrencia de un evento, y de la relación entre un sistema global de acontecimientos y los eventos individuales que lo constituyen Shannon estableció también una definición matemáticamente rigurosa de la información media esperada de una cadena de eventos producidos por un sistema o fuente de información.

C.1.- Sistema generador de eventos o fuente de información:

Una fuente de información F es todo conjunto de n posibles eventos independientes x_i ($i = 1, 2, \dots, n$) capaces de ocurrir y de sus respectivas probabilidades de ocurrencia $P(x_i) = p_i$, de nuevo bien definidas según Kolmogorov¹⁸:

$$0 \leq p_i \leq 1 \quad \text{y} \quad \sum_{i=1}^n p_i = 1 \quad \text{para } i = 1, 2, \dots, n$$

Cualquier sistema de la naturaleza emisor de símbolos o productor de una serie de acontecimientos constituye una fuente de información en el sentido que acabamos de definir. Para esta forma de ver las cosas, la distribución de probabilidades es una expresión de la esencia y la estructura de todo el sistema, y por ello a esas probabilidades se les denominan "probabilidades estructurales" del sistema.

C.2.- Información media (esperada) de un sistema: Entropía de Shannon:

Una fuente de información F , puede caracterizarse mediante la medida de la información que esperamos obtener de la ocurrencia de cada uno de sus acontecimientos. La información esperada de cada fuente o sistema es la media (esperanza matemática) de las informaciones proporcionadas por cada uno de los acontecimientos que puede generar, una vez que estos ocurran.

Según hemos definido antes, la información proporcionada por la ocurrencia de cada acontecimiento x_i generado por la fuente o sistema F es:

$$I_i = I(p_i) = -\log_2 p_i \text{ bits}$$

Se define como información media (por evento producido) de la fuente F o **entropía de Shannon** $H(F)$ a la media (esperanza matemática) de la información obtenida de los acontecimientos generados por la fuente:

$$H(F) = I_1 \cdot p_1 + I_2 \cdot p_2 + \dots + I_n \cdot p_n = \sum_{i=1}^n I_i \cdot p_i$$

es decir,

$$\begin{aligned} H(F) &= -\sum_{i=1}^n p_i \log_2 p_i \text{ bits por evento} = \\ &= \sum_{i=1}^n p_i \log_2 (1/p_i) \text{ bits por evento} \end{aligned}$$

Se usa la convención de que $0 \log_2 0 = 0$, que se justifica fácilmente por el argumento de continuidad, ya que $x \log_2 x \rightarrow 0$ cuando $x \rightarrow 0$. Así añadir términos con probabilidad cero no cambia la entropía del sistema. Nótese también que la entropía de un sistema *no depende* de los valores que tome cada uno de los acontecimientos, sino sólo de las probabilidades estructurales de los eventos individuales.

El **segundo teorema** de Shannon demuestra que la entropía $H(F)$ de una fuente de información F satisface la condición

$$0 \leq H(F) \leq \log_2 n ,$$

donde $H(F) = 0$ si, y sólo si, existe un i tal que $p_i = 1$, y donde el valor máximo de la entropía del sistema se produce cuando $H(F) = \log_2 n$, situación que se produce si, y sólo si, $p_i = 1/n$ para todo $i = 1, 2, \dots, n$, es decir en la situación de equiprobabilidad de los eventos que pueden ocurrir. Así, un sistema que puede producir dos eventos equiprobales (lanzar una moneda perfecta) tiene una entropía de 1 bit por tirada, su máximo posible, pero se puede calcular que si la moneda está sesgada ($p_{\text{caras}} = 0.8$, $p_{\text{cruces}} = 0.2$) la entropía del sistema es entonces de 0.72 bits por tirada [FIGURA IV-1]. La entropía del sistema formado por un dado perfecto es de 2.58 bits por tirada. La demostración de este segundo teorema, aunque es simple (se basa en una sencilla propiedad de la función logarítmica) y se deriva del primer teorema -la entropía es siempre positiva por ser una media de cantidades positivas-, excede también el objetivo de este trabajo y puede consultarse en la literatura citada^{245,246}.

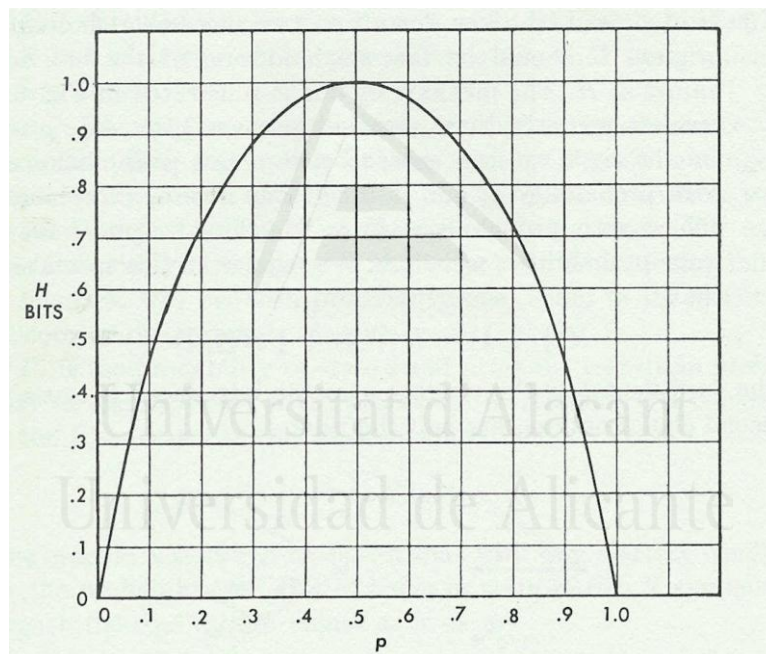


FIGURA IV-1: En un sistema indeterminista como lanzar una moneda, siendo p la probabilidad de cara y $q = 1 - p$ la de cruz, la curva que relaciona la Entropía del sistema (en bits por tirada) respecto a los valores de p es simétrica: tiene un valor máximo de 1 cuando $p = q = 0.5$ (máxima indeterminación) y es 0 cuando p (o q) valen 0 o 1 (sistema determinista). Tomado de Shannon CE⁵⁸

C.3.- Información, entropía, desorden e indeterminismo del sistema:

Todo ello confiere un profundo significado al concepto de entropía de un sistema generador de información: la entropía es una magnitud que *mide el grado indeterminismo o aleatoriedad del sistema*, su grado de complejidad en relación a la variabilidad aleatoria de sus componentes. Un sistema determinista tiene una entropía igual a cero; cuanto más indeterminista es el sistema (mayor número de

eventos posibles equiprobables entre ellos) mayor es su entropía. No es de extrañar, pues, que, por ejemplo, en biología autores como R. Margalef o J. W. McArthur hayan propuesto con éxito utilizar la entropía de Shannon como medida de la diversidad biológica de un ecosistema: la biodiversidad del Amazonas estaría más próxima a la del arca de Noé que a la de una granja... La entropía se usa también en lingüística -para cuantificar en que medida una lengua es predecible y, en consecuencia, su redundancia-, en criptoanálisis -para intentar revelar algoritmos criptográficos- y en ingeniería informática -en el diseño de programas para el reconocimiento del habla-.

El uso del término *entropía* de un sistema tiene sus orígenes en la ciencia física, y acepta dos definiciones diferentes, según el sistema pertenezca al mundo microscópico o al macroscópico. En el mundo macroscópico, el concepto fue introducido de forma intuitiva en 1824 por Sadi Carnot en sus "*Reflexiones acerca de la fuerza motriz del fuego*", y formalizado en 1865 por Rudolf Clausius para enunciar la segunda ley de la termodinámica. En este contexto, la entropía permite diferenciar la parte de energía que puede convertirse en trabajo de aquella otra que se traduce, por ejemplo, en calentamiento, y desarrolló un importante papel en la mejora del rendimiento de las primeras máquinas de vapor. A escala microscópica, en la física estadística "entropía" representa la *capacidad* que tiene un sistema de adquirir un gran número de estados, caracterizados por la energía cinética, potencial y electromagnética de sus componentes. Max Planck llegó, a partir de los trabajos de Ludwig Boltzmann, a la expresión $S = k \log W$ en la que W representa el número de estados. En el caso de un gas perfecto, en el que las moléculas se consideran esferas, la distribución de estas en el espacio basta para determinar la entropía del sistema. Es por ello que, en física, la entropía se describe a menudo como una medida del "desorden" o "desorganización" de un sistema.

En ciertos casos, la entropía se expresa como una suma de logaritmos: un sumatorio de $x_i \log x_i$ siendo x_i el número de estados individuales en los que puede encontrarse una partícula dada. Por analogía formal, el término fue introducido en la teoría matemática de la información por su fundador Claude E Shannon. Cuando Shannon se dio cuenta de la importancia de su expresión para $H(F)$, consultó al gran matemático John von Neumann buscando que le ayudara a encontrarle un nombre adecuado. Cuenta la tradición que la respuesta de von Neumann fue como sigue: "debes llamarla *entropía* por dos razones: primero esta misma función se usa en termodinámica con ese nombre, pero segundo, y más importante, la mayoría de la gente no sabe que es realmente la entropía, iiy si utilizas la palabra *entropía* en tus argumentos iisiempre saldrás victorioso!!".

D. - Entropía relativa o Divergencia logarítmica de Kullback-Leibler:

Para la teoría de probabilidades, la *entropía* $H(X)$ de una variable aleatoria discreta X con una función de distribución de probabilidad $p(x_i)$ se define como:

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \text{ bits}$$

y es una medida de la aleatoriedad o indeterminismo de la variable: de la cantidad de información que, como media, nos aporta el sistema cuando se produce la ocurrencia del evento incierto. David Applebaum ha demostrado el teorema de unicidad de la entropía: que la entropía es la única medida posible de la aleatoriedad o indeterminismo de la variable²⁴⁶.

Así, por ejemplo, podemos calcular la entropía de una variable discreta que tenga una distribución uniforme sobre 32 eventos (por ejemplo hacer girar una ruleta de 32 números). La entropía de este "sistema aleatorio" será:

$$H(X) = - \sum_{i=1}^{32} p(x_i) \log_2 p(x_i) = - \sum_{i=1}^{32} (1/32) \log_2(1/32) = \log_2 32 = 5 \text{ bits}$$

Pero supongamos ahora que asistimos a una carrera de caballos en la que participan ocho corceles, cada uno con una probabilidad de ganar muy diferente. Así, por ejemplo, estimadas con un sistema de apuestas, las probabilidades de ganar son: (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64 y 1/64). Pues bien, podemos calcular la entropía de este sistema aleatorio que constituye nuestra carrera de caballos:

$$H(X) = - 1/2 \log_2 1/2 - 1/4 \log_2 1/4 - 1/8 \log_2 1/8 - 1/16 \log_2 1/16 - 4 \cdot 1/64 \log_2 1/64 = \\ = 2 \text{ bits}$$

La carrera es un sistema mucho más determinista que la ruleta: es más difícil adivinar que número de la ruleta va a salir que el caballo que va a ganar. Es más arriesgado jugar a esta ruleta que apostar a la carrera.

Sea X una variable aleatoria discreta que puede distribuirse con dos funciones distintas de distribución de probabilidad $p(x_i)$ y $q(x_i)$. Kullback y Leibler^{90,89} han utilizado la entropía de Shannon para definir una **distancia de indeterminismo** entre ambas distribuciones de probabilidad, entre ambos sistemas aleatorios. Así definen la **entropía relativa** (divergencia logarítmica Kullback-Leibler) como:

$$D(p(x)||q(x)) = \sum_{i=1}^n p(x_i) \log_2 [p(x_i)/q(x_i)] \text{ bits}$$

es decir, desde el punto de vista estadístico, la esperanza del logaritmo de la razón de verosimilitudes de cada valor de la variable. En esta definición se utiliza la convención (basada en argumentos de continuidad) de que $0 \log_2(0/q) = 0$, y de que $p \log_2(p/0) = \infty$.

D.1.- Discrepancia intrínseca:

La entropía relativa $D(p||q)$ suele interpretarse como una "distancia" entre distribuciones de probabilidad, una medida de la información que se pierde cuando se asume que la distribución es $q(x)$, si la verdadera distribución es $p(x)$. Así, si conociéramos la verdadera distribución $p(x)$ podríamos medir su entropía $H(p)$, pero si asumimos $q(x)$, necesitaremos como media $H(q) + D(p||q)$ para conocer la verdadera entropía de la variable aleatoria. Esto lo mismo que decir que la entropía relativa mide el **indeterminismo o aleatoriedad que estamos introduciendo** en el sistema si se asume que la distribución es $q(x)$, cuando la verdadera distribución es $p(x)$.

Pero, estrictamente hablando, por sus propiedades matemáticas, la entropía no es una distancia estricta. Así, se puede demostrar que $D(p||q) \geq 0$ con igualdad si, y sólo si, $p(x) = q(x)$, pero, en general $D(p||q) \neq D(q||p)$ y con esta asimetría no se satisface la desigualdad triangular. Por ello en la literatura se ha propuesto expresar este mismo concepto con otras medidas con propiedades más adecuadas al concepto de "distancia" entre distribuciones, como por ejemplo la Divergencia de Jeffreys = $D(p||q) + D(q||p)$ bits o la Discrepancia intrínseca²⁴⁷ = $\text{Mín}[D(p||q), D(q||p)]$ bits de Bernardo.



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

ANEXO V: EL ÍNDICE "PEDIATRIC RISK OF MORTALITY" (PRISM):

	< 1 año	Todos	> 1 año	Puntuación
TAS (mm Hg)	130-160		150-200	2
	55-65		65-75	2
	> 160		> 200	6
	40-54		50-64	6
	< 40		< 50	7
TAD (mm Hg)		> 110		6
Frecuencia cardíaca (lpm)	> 160		> 150	4
	< 90		< 80	4
Frecuencia respiratoria (rpm)	61-90		51-70	1
	> 90		> 70	5
	Apnea		Apnea	5
PaO ₂ /FiO ₂		200-300		2
		< 200		3
PaCO ₂ (torr)		51-65		1
		> 65		5
Glasgow		< 8		6
Reacción pupilar		anisocoria/midriasis fijas y midriáticas		4 10
TP/TPT		> 1,5 veces/control		2
Bilirrubina total (mg/dl)		> 3,5 en > 1 mes		6
Potasio (mEq/L)		3-3,5		1
		6,5-7,5		1
		< 3		5
		> 7,5		5
Calcio (mg/dl)		7-8		2
		12-15		2
		< 7		6
		> 15		6
Glucosa (mg/dl)		40-60		4
		250-400		4
		< 40		8
		> 400		8
Bicarbonato (mEq/L)		< 16		3
		> 32		3

TP: Tiempo de protrombina. TPT: Tiempo parcial de tromboplastina.

CÁLCULO DEL LOGIT PRISM:

$$(0,207 \times \text{PRISM}) - (0,005 \times \text{edad en meses}) - (0,433 \times \text{cirugía previa [Sí = 1, No = 0]}) - 4,782$$

PROBABILIDAD DE MUERTE: $(e^{\text{Logit}} / 1 + e^{\text{Logit}}) \times 100 = \% \text{ de mortalidad}$



Universitat d'Alacant
Universidad de Alicante

ANEXO VI: REDES NEURONALES ARTIFICIALES

En el afán de potenciar los sistemas de decisión, hoy en día es posible construir computadoras que realicen una amplia variedad de tareas bien definidas con una celeridad y seguridad no permitidas para los humanos. A pesar de ello existen problemas para los cuales estas máquinas no proporcionan una solución aceptable; son tareas generalmente difíciles de concretar y que usualmente requieren gran cantidad de operaciones, como puede ser el reconocimiento de imágenes. Ningún sistema de computación en visión puede rivalizar con la capacidad humana para reconocer imágenes. La explicación a este hecho está en que la información que proviene del mundo real es masiva, redundante e imprecisa, mientras que el computador está orientado a trabajar con datos precisos y de forma secuencial.

Por tanto, la resolución de ciertos problemas necesita de otro tipo de procesadores diferentes a los computadores digitales clásicos. En estos casos las redes neuronales son una alternativa eficaz gracias a sus características de tratamiento no lineal de los datos, tolerancia a fallos y trabajo en paralelo. Estos sistemas son una de las herramientas más utilizadas en la actualidad en el tratamiento de la información. Han despertado tanta expectación no sólo porque exhiben interesantes propiedades sino también porque proporcionan un marco de estudio a los métodos de procesado y tipos de información empleados por el sistema nervioso.

A.-Introducción a las Redes Neuronales Artificiales (RNA):

La aparición de las RNA ha estado vinculada a las investigaciones del sistema nervioso biológico, en especial el humano. El elemento de procesado del sistema nervioso es la neurona y, por tanto, su fisonomía es una de las claves que permiten que el cerebro exhiba esas propiedades tan interesantes. En 1888 Ramón y Cajal, gracias a sus aportaciones a la teoría reticular, demostró que el cerebro está compuesto, en realidad, por una red de células (neuronas), compuestas de axones, somas y dendritas. Más tarde postuló que las neuronas, como células altamente especializadas, determinan la dirección de transmisión de la información dentro del sistema nervioso. De esta forma concibió el cerebro como un órgano altamente complejo, paralelo y jerarquizado.

Alan Turing, en 1936, fue el primero en estudiar el cerebro como una forma de ver el mundo de la computación. Sin embargo no fue hasta 1943 cuando Warren McCulloch, neurofisiólogo, y Walter Pitts, matemático, sugirieron una teoría acerca de la forma de trabajar de las neuronas y constituyeron los fundamentos de la computación neuronal. Confeccionaron el primer prototipo de neurona artificial, un modelo muy simple pero que contenía todas las características básicas de las neuronas. Ambos demostraron que mediante combinaciones de sus neuronas se podía construir cualquier función lógica.

Una de las primeras aportaciones importantes en el aprendizaje de los sistemas biológicos fue proporcionada por Donald Hebb en 1949. Poco tiempo antes se había descubierto que la transmisión de información, dentro del sistema nervioso, tiene lugar en las uniones o sinapsis entre neuronas. Hebb propuso que el reforzamiento de la sinapsis entre dos neuronas era proporcional a la actividad de la conexión: "... cuando un axón presináptico causa la activación de cierta neurona postsináptica, la eficacia de la sinapsis que las relaciona se refuerza" ²⁴⁸.

El trabajo experimental posterior ha confirmado en parte esta teoría demostrando la presencia de este tipo de aprendizaje en la neurona biológica aunque en coexistencia con otros esquemas. Esta regla de aprendizaje, en principio obtenida en el campo biológico, ha resultado muy útil para resolver determinados problemas tecnológicos.

Fue necesario esperar hasta 1957 para que Frank Rosenblatt construyera el Perceptrón, la primera red neuronal artificial con proyección comercial, basándose en la estructura y funcionamiento de las neuronas receptoras de la retina. Su primera aplicación fue la clasificación de patrones visuales. El Perceptrón simple [FIGURA VI-1] estaba constituido por una sola neurona que, a diferencia de la neurona de McCulloch que tenía arquitectura fija, permitía adaptarse a diferentes tareas modificando las conexiones de las entradas gracias a un algoritmo ideado por el propio Rosenblatt²⁴⁹.

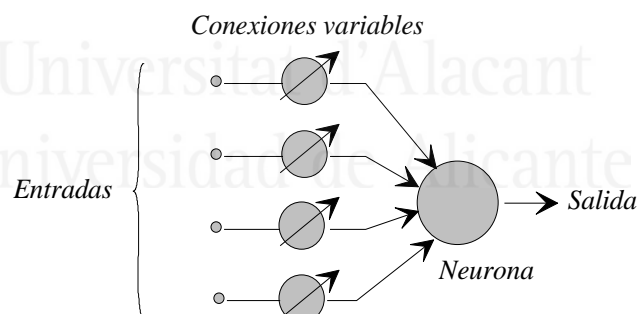


Figura VI-1. Esquema del perceptrón.

En 1959, Bernard Widrow y Marcial Hoff desarrollaron una variante del Perceptrón al que dieron el nombre de ADALINE (*AD*aptative *L*ineal *N*euron). Modificaba sus conexiones en función de la tarea a realizar mediante un nuevo algoritmo al que llamaron LMS (*L*east *M*ean *S*quare)²⁵⁰. Su enorme potencialidad pronto se aprovechó en diferentes ambientes, especialmente en el campo de la comunicación donde se aplicó como cancelador de ecos o ecualizador de canal.

En los años siguientes a esos descubrimientos se realizaron grandes avances en el campo de las redes neuronales y la disciplina creció rápidamente hasta que, en

1969, Marvin Minsky y Seymour Papert publicaron *Perceptrons*, un famoso libro en el que demostraban la incapacidad del perceptrón simple y la adalina para resolver problemas de clasificación que no fueran separables linealmente y donde se hacía la conjetura que la extensión a varias capas de neuronas no sería de utilidad. A partir de este libro, surgieron numerosas críticas en contra de las redes que frenaron el crecimiento de las investigaciones sobre las redes neuronales.

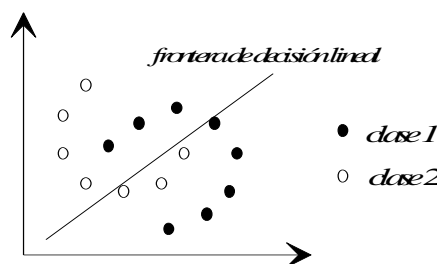


FIGURA VI-2. Grupos no separables mediante el perceptrón lineal.

No se consiguió resolver el problema de la extensión del perceptrón simple a uno de varias capas (multicapa) y, por tanto, solucionar el problema de la separabilidad lineal [FIGURA VI-2] hasta que Werbos, en 1974, publicó un algoritmo que permitía ajustar las conexiones de las neuronas en las redes multicapa (con conexiones hacia delante). El algoritmo, muy popular en el entorno de las RNA, es conocido como *backpropagation*. Desafortunadamente el trabajo de Werbos permaneció desconocido en la comunidad científica. En 1982 Parker redescubrió la técnica y la publicó en el Instituto de Tecnología de Massachussets. No mucho después Rumelhart, Hinton y Williams la volvieron a descubrir y la popularizaron. Es uno de los mayores avances en redes neuronales puesto que abrió el camino para lo que más tarde ha sido la red neuronal más aplicada, el perceptrón multicapa.

Otros grandes avances en el campo de las redes neuronales han llegado al intentar emular el funcionamiento y particularidades del cerebro. Una de las singularidades más relevantes de la memoria humana es la habilidad que tiene para aprender nuevos conceptos sin por ello olvidar los aprendidos en el pasado. Sin embargo, muchos de los modelos de redes neuronales artificiales pierden gran parte de la información aprendida cuando se les entrena por segunda vez. En 1986, con la intención de resolver este problema, que se ha dado en llamar el dilema de la estabilidad y plasticidad en el aprendizaje, Stephen Grossberg y G. Carpenter presentaron su red ART (*Adaptive Resonant Theory*)²⁵¹. La idea consiste en agrupar la información de entrada en función de la similitud que presenta frente a prototipos creados por la red, creando nuevas clases si el grado de semejanza no supera cierto umbral con lo que se evita destruir categorías anteriormente creadas.

Existen muchas evidencias sobre la organización de las neuronas de forma que las informaciones captadas del entorno a través de los órganos sensoriales se representan internamente en forma de mapas bidimensionales. Es probable que parte de ella se origine mediante el aprendizaje. Por tanto, el cerebro podría poseer la capacidad inherente de formar mapas topológicos de las informaciones recibidas del exterior. Teuvo Kohonen presentó en 1982 un modelo de red neuronal, denominado SOM (*Self Organization Maps*), con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro²⁵².

Otra de las redes que más repercusión ha tenido es el Neocognitrón. Es una red diseñada por Kunihiro Fukushima, en 1982, para tareas de procesado de imágenes tales como reconocimiento de caracteres. Las características que hacen esta red única es una conectividad selectiva a lo largo de sus capas jerarquizadas. Las características de la imagen de bajo nivel son detectadas en las primeras capas y combinadas para formar objetos más generales en las siguientes capas. Se ha demostrado que es capaz de reconocer objetos independientemente de su localización en una imagen, de las deformaciones o de oclusiones parciales del objeto²⁵³.

Otro de los principales responsables del desarrollo que ha experimentado el campo de la computación neuronal ha sido John Hopfield quien construyó un modelo de red, en 1982, con el número suficiente de simplificaciones como para poder extraer analíticamente información del sistema. Más recientemente Bart Kosko extendió algunas de las ideas de la red ART y la de Hopfield para desarrollar su BAM (*Adaptative Bidireccional Associative Memory*), un modelo de red que emplea diferentes reglas de aprendizaje.

Una alternativa a los modelos neuronales descritos son las redes estocásticas cuya salida se obtiene de forma probabilística y con mecanismos de aprendizaje también estocásticos, basados en la idea de seleccionar de forma aleatoria valores para los pesos de las conexiones y comprobar el efecto en el rendimiento de la red. La cualidad más interesante es la capacidad para escapar de los mínimos locales gracias a su comportamiento aleatorio. La red más conocida que responde a este funcionamiento estocástico es la denominada máquina de Boltzman, ideada en 1984 por Hinton, Sejnowski y Ackley. La máquina de Cauchy, concebida por Szu (1986), es una versión mejorada de la máquina de Boltzman con una arquitectura y funcionamiento idénticos excepto en lo que concierne a la función de probabilidad y a la función de temperatura que establece el plan de templado o enfriamiento de la red²⁵¹.

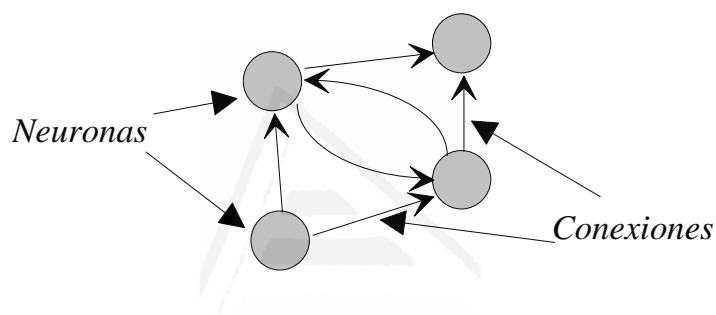
B. - Definición:

Las RNA son un intento de modelizar las capacidades de procesado de información del sistema nervioso. Pueden ser consideradas como una aproximación más al problema de la computación y, consecuentemente, realizan las acciones de

almacenamiento, transporte y procesado. Las neuronas son, en las RNA, sus elementos más pequeños, las que procesan y transportan la información, mientras que el almacenamiento se produce en las conexiones o sinapsis entre ellas. Una definición formal podría ser²⁵⁴ [FIGURA VI-3]:

Una red neuronal es un procesador de cálculo distribuido que tiene una tendencia a almacenar conocimiento experimental existiendo la posibilidad de usar este conocimiento. Este procesador se parece al cerebro en dos aspectos: a) El conocimiento es adquirido a través de un proceso de aprendizaje regido por un algoritmo de aprendizaje (*learning algorithm*); y b) .Las conexiones entre los elementos base (neuronas) - conocidos como pesos sinápticos-, son usados para el amacenamiento de este conocimiento

FIGURA VI-3: Representación gráfica de una red neuronal artificial.



C.- RNA en Ciencias de la Salud:

C.1.- Potencial de las RNA en medicina:

Las redes neuronales artificiales han sido aplicadas, con notable éxito, en problemas de clasificación, modelización, procesado de señales y predicción de series temporales en campos tan dispersos como la ingeniería, la economía o las telecomunicaciones. Expertos en otras materias han reconocido el potencial de la nueva técnica y la están incorporando a sus trabajos.

Sin embargo, en el ámbito de las ciencias de la salud, se prefieren los métodos estadísticos clásicos²⁵⁵, tales como la regresión logística, para el análisis de datos. Así, por ejemplo, en un estudio realizado sobre los 311 artículos publicados en *New England Journal of Medicine* entre Enero de 2004 y Junio de 2005, el 51% utilizó la regresión logística en su estudio, quedando detrás sólo de otras técnicas estadísticas convencionales como el análisis de tablas de contingencia (prueba χ -cuadrado y/o prueba exacta de Fisher) y el análisis de supervivencia²⁵⁶. La base matemática sobre la que descansa esta metodología, su relativa sencillez y, especialmente, su extendida popularidad entre los facultativos, justifican para muchos investigadores, seguir utilizándola. Una cuestión importante sobre las técnicas estadísticas habitualmente empleadas es que son lineales en los

parámetros, es decir, se presupone que las relaciones entre las covariables del problema no revisten excesiva complejidad.

En cambio, en las disciplinas de las ciencias de la salud se realizan estudios en los que intervienen multitud de factores para los cuales, en muchas ocasiones, se obvian las interacciones o, simplemente, no se evalúan con suficiente rigor. Existen infinidad de variables para describir la situación y el contexto de un paciente, tantas que, hoy en día, se presupone que la complejidad que surge de las interacciones entre los diferentes estados clínicos genera un contexto cuya comprensión se vuelve intratable²⁵⁷. La simplificación del problema, efectivamente, facilita el análisis pero, por contrapartida, conlleva una pérdida de generalización y de precisión en el estudio.

Las redes neuronales son capaces de captar los matices que se escapan a los métodos estadísticos más simples. En efecto, hay evidencias de los buenos resultados proporcionados en tareas bien definidas donde las interacciones entre las covariables del problema son significativas²⁵⁷. Se ha establecido que las redes neuronales son equivalentes a técnicas estadísticas paramétricas y no-paramétricas²⁵⁸. Sin embargo, las redes realmente ocupan un lugar intermedio y privilegiado: actúan como técnicas semi-paramétricas -es decir, más flexibles que los métodos paramétricos-, pero requieren menos parámetros que los métodos no-paramétricos²⁵⁹. Otras ventajas son inherentes a su constitución y fundamentos neurobiológicos²⁵⁴:

- Tratamiento no lineal de la información proporcionado por la interconexión de elementos simples de procesamiento no lineal (neurona).
- Capacidad de establecer relaciones entrada-salida a través de un proceso de aprendizaje.
- Aprendizaje adaptativo que les permite a llevar a cabo ciertas tareas mediante un entrenamiento con ejemplos ilustrativos. No es necesario tener modelos a priori ni se necesita especificar funciones de distribución de probabilidad.
- Robustez o tolerancia a fallos dado que almacenan la información aprendida de forma distribuida en las conexiones entre neuronas.
- Uniformidad de análisis y diseño proporcionado por teorías conjuntas que describen los diferentes algoritmos y aplicaciones.

En general, el uso de las redes neuronales está justificado en problemas donde se pueda aprovechar su potencial para el análisis no lineal de la información, su función como memoria asociativa distribuida como forma de evitar las dificultades en la adquisición de conocimiento experto, tolerancia al ruido, proporcionada por su arquitectura inherentemente paralela, y su adaptabilidad para acomodarse a nuevas manifestaciones de la enfermedad. Estas características hacen a las redes neuronales preferibles a otros métodos matemáticos en problemas para los cuales⁶⁸:

- No es posible encontrar un conjunto de reglas sistemáticas que describan completamente el problema.
- Se dispone de una cantidad razonable de ejemplos representativos del problema.
- Hay que trabajar con datos imprecisos o incoherentes.
- Se tiene un gran número de variables que definen el problema (alta dimensionalidad del problema).
- Las condiciones del problema son cambiantes.

Estas condiciones reflejan, fielmente, el entorno de trabajo de los facultativos clínicos. La toma de decisiones clínicas se realiza bajo condiciones de incertidumbre en la información, unas veces imprecisa y otras veces incoherente²⁶⁰. El análisis de situaciones complejas, generalmente, involucra un gran número de factores que, desafortunadamente, no siempre se sabe a ciencia cierta como interpretarlos para tomar una decisión. Las redes neuronales, aprendiendo de los ejemplos reales, son capaces de procesar grandes cantidades de datos para extraer características relevantes y útiles de los datos. De esta forma se reduce el nivel de complejidad del problema y se eleva el nivel de abstracción de la información, más próxima al médico y, por tanto, es más fácil de asimilar y aprovechar para determinar la acción a realizar.

C.2.- Aplicaciones en la medicina clínica.

El campo de la medicina no ha sido inmune a la tendencia de aplicar las redes neuronales. El aliciente ha sido la mejora de resultados en problemas complejos, generalmente tratados con técnicas estadísticas clásicas. La búsqueda de un mejor tratamiento de los datos o la explotación de grandes bases de datos médicas para la extracción de evidencias ha justificado el incremento en la aplicación de las redes neuronales. Entre 1993 y 2000, en una búsqueda bibliográfica realizada en MEDLINE²⁶⁰, había 3101 artículos relacionados con las redes neuronales en una clara evolución creciente [FIGURA VI-4]:

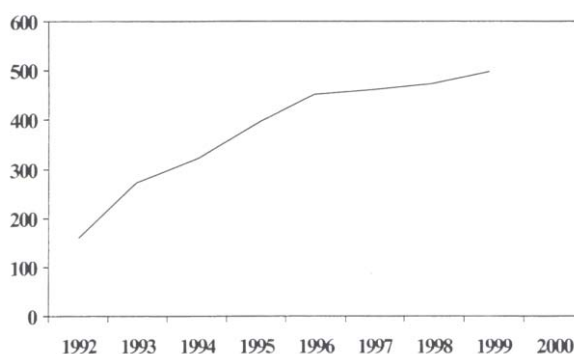


FIGURA VI-4: Artículos encontrados en MEDLINE que aplicaban RNA.

El potencial de las redes continúa creciendo cubriendo no sólo áreas de predicción de estados clínicos, sino también evaluando la posibilidad de enfermedad desde datos, o desde bases de datos complejas. La estructura y

plasticidad de las redes neuronales permitirán aplicarla a otros problemas insospechados.

En una revisión comparativa sobre aplicaciones de las redes en ensayos aleatorizados y controlados^{257,259,261} se realizaron comparaciones entre los resultados obtenidos por las redes neuronales y por las aproximaciones anteriormente empleadas. Principalmente la comparación versó entre las técnicas más usadas por ambas partes, el perceptrón multicapa y la regresión logística con resultados equiparables o superiores para la red neuronal en la práctica totalidad de los estudios. Además se han realizado comparaciones entre las redes neuronales y otras metodologías. El más conocido es el proyecto Statlog, en el cual se realiza una extensa comparación de las redes neuronales, con la estadística tradicional y con los árboles de decisión sobre 22 conjuntos de datos, 3 de ellos médicos. En él se muestra que la red obtiene mejores resultados en sólo un conjunto de datos²⁵⁵, pero se admite que las redes proporcionaron el menor error y, por tanto, generaron los mejores modelos predictivos en la práctica totalidad de los casos.

C.3.- Ayuda al diagnóstico:

La diagnosis es el proceso de determinar la condición de una enfermedad examinando la naturaleza y circunstancias del paciente. Es un proceso de clasificación. Las características de las RNA las hacen idóneas para tratar problemas de clasificación. Las reglas de aprendizaje basadas en la minimización del error cuadrático, proporcionan una interpretación estocástica de las RNA como estimadores de una distribución de probabilidad condicionada, directamente y sin asunciones sobre la estructura probabilística de los datos. Si unimos este hecho con la capacidad que tienen las RNA para aproximar cualquier función con un grado de precisión arbitraria, entonces resulta que, dados los suficientes datos, recursos computacionales y tiempo, es posible, usando una red de propagación hacia delante, estimar mediante los valores de salida, las probabilidades a posteriori de pertenencia a una clase utilizando el Teorema de Bayes²⁶².

Prácticamente, en la totalidad de los campos de la medicina se han aplicado las redes neuronales con el objetivo de proporcionar soporte al diagnóstico. En Cardiología, como asistencia en la auscultación²⁶³, pero también para la detección de isquemias²⁶⁴ y la prevención, mediante el diagnóstico precoz, del infarto de miocardio mediante datos del historial del paciente²⁶⁵ o a partir de marcadores biométricos²⁶⁶ y de la muerte súbita después del infarto²⁶⁷. En Hemato-Oncología, se han utilizado para la detección precoz, basándose en parámetros biométricos, del cáncer de próstata en hombres obteniendo mejores resultados que la regresión logística²⁶⁸, y también se han utilizado para diagnosticar cáncer de mama^{269,270}, como ayuda al diagnóstico de anemias²⁷¹, para identificación de pacientes con bajo riesgo de reaparición del cáncer²⁷², y en estudios sobre supervivencia en pacientes tratados con cánceres situados en la cabeza y cuello²⁷³. En Urología, se han utilizado para la ayuda al diagnóstico en la determinación de la infertilidad

masculina mediante biopsias de testículos y para pronosticar metástasis y mortalidad de pacientes con cáncer renal, mejorando resultados sobre el análisis discriminante lineal y cuadrático²⁷⁴ o aplicadas al tratamiento endoscópico del reflujo vesicoureteral en niños²⁷⁵. En Neurología, se han usado para diferenciar entre la enfermedad de Alzheimer y demencia vascular²⁷⁶, y la empresa Oxford Biosignals BioSleep (BioSleep) [FIGURA VI-5] utiliza las redes neuronales para analizar continuamente la estructura del sueño para detectar *microarousals* ("microdespertares") a partir de un único canal de EEG.

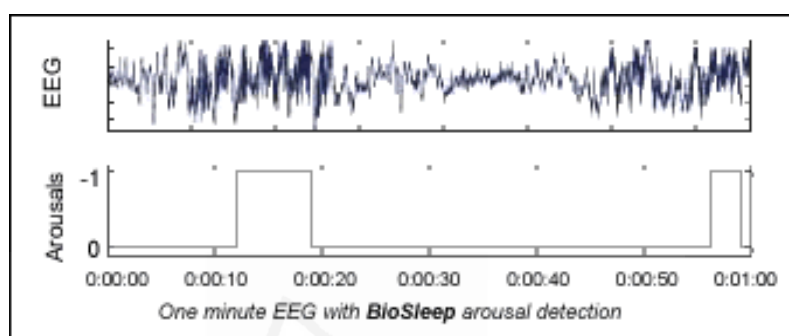


FIGURA VI-5: Sistema de detección de microarousals BioSleep. <http://www.oxford-biosignals.com/>.

Un campo en el que las RNA se están aplicando con éxito es el de la Farmacocinética. La predicción de concentraciones de fármacos en sangre es un problema complejo ocasionado por la fuerte dependencia con el metabolismo del paciente. Las redes neuronales, a partir del seguimiento o monitorización, capturan las relaciones entre los niveles plasmáticos de fármacos y las características de los pacientes, y modelizan así la cinética de los fármacos en el organismo pudiendo ser utilizadas para realizar predicciones personalizadas²⁷⁷. Esta cualidad es importante para fármacos que tienen estrechos rangos terapéuticos, en especial, aquellos que presentan toxicidad al superar ciertas concentraciones plasmáticas. Así, se han aplicado con éxito a la predicción de concentraciones plasmáticas de Gentamicina en pacientes afectados con infecciones serias²⁷⁸, y a la predicción de los niveles de Tacrolimus en sangre posteriores a trasplante de hígado²⁷⁹.

También se aplican las RNA en el análisis de las muestras de laboratorio, dadas las ventajas de estimación sin suposiciones de modelos, generalización y capacidad para procesar datos de manera no lineal. Por ejemplo, en Microbiología, se utilizan redes neuronales para analizar espectros de masas obtenidos mediante pirólisis (separación de sustancias químicas usando calor) para identificar bacterias. Ejemplos son la identificación de la bacteria de la tuberculosis²⁸⁰ o la detección de estreptomicetos²⁸¹.

En Anestesiología, los médicos requieren experiencia para evaluar las diferentes señales de monitorización de los pacientes. Las RNA pueden ayudar a

los médicos al tratar esa información más rápidamente. Así, se ha comprobado que se puede reducir el tiempo de respuesta desde los 45 segundos de media de los clínicos a 17 proporcionados por las redes²⁶¹. En Cuidados Intensivos, se han aplicado redes neuronales para clasificar el nivel de sedación a partir del análisis del electroencefalograma²⁸², o para predecir la hemorragia intracraneal en pacientes neonatos²⁸³.

C.4.- Perspectiva futura:

Los buenos resultados de las redes como potentes procesadores de información clínica aseguran un uso generalizado con el paso del tiempo. Sin embargo, hay pocas evidencias, por parte de los clínicos, que los prototipos obtenidos en las investigaciones sean desarrollados más allá. La falta de rigor de algunos trabajos realizados con redes neuronales, en los que no se detalla la población utilizada en el estudio, no se describe el proceso de desarrollo de los modelos o no se realizan comparaciones con otras aproximaciones están generando grandes críticas que desacreditan el valor de las redes neuronales. Para que las redes neuronales sean realmente aceptadas a nivel clínico, es esencial que se realicen estudios rigurosos sobre su rendimiento como métodos de diagnóstico²⁵⁷.

Las redes neuronales, a pesar de proporcionar mejores resultados que otras aproximaciones, tienen multitud de detractores porque las imaginan como *cajas negras*. La falta de comprensión originada por la dificultad para explicar el resultado proporcionado por la red en términos simples es, en muchas ocasiones, motivo suficiente para desacreditarlas. Pero, a nuestro entender, las RNA están llamadas a desempeñar un importante papel en la práctica clínica, siempre y cuando se consiga un alto grado de integración con los los sistemas de información de soporte a la decisión en los escenarios clínicos donde sus capacidades predictivas y de diagnóstico demuestren ser realmente relevantes.

D.- Estructura de una Red Neuronal Artificial:

Las RNA, al igual que las biológicas, están formadas por nodos elementales de proceso interconectados. Así pues la estructura de la red estará definida si se describe la forma de operar de los nodos y la manera en la que éstos se relacionan.

Primero se mostrará un modelo general de neurona artificial. Es un modelo demasiado amplio para lo que se utiliza en las principales redes conocidas pero tiene la peculiaridad de explicar las propiedades esenciales de las neuronas biológicas a la vez que nos permite agrupar, bajo un mismo modelo, los diferentes tipos de neuronas utilizadas en aplicaciones prácticas.

D.1.- Neurona:

Se denomina neurona artificial a un dispositivo simple de cálculo que, a partir de un vector de entrada procedente del exterior o de otras neuronas, proporciona una única respuesta o salida. Generalmente se pueden encontrar tres tipos de neuronas:

- 1) Las neuronas de entrada que reciben estímulos externos y adquieren la información.
- 2) La información se transmite a otras neuronas que no tienen relación directa con la información de entrada también denominada neuronas ocultas.
- 3) Al final la información llega a las neuronas de salida que presentan la respuesta de la red.

Los elementos que componen la estructura de neurona general más extendida son los siguientes⁶⁸ [FIGURA VI-6]:

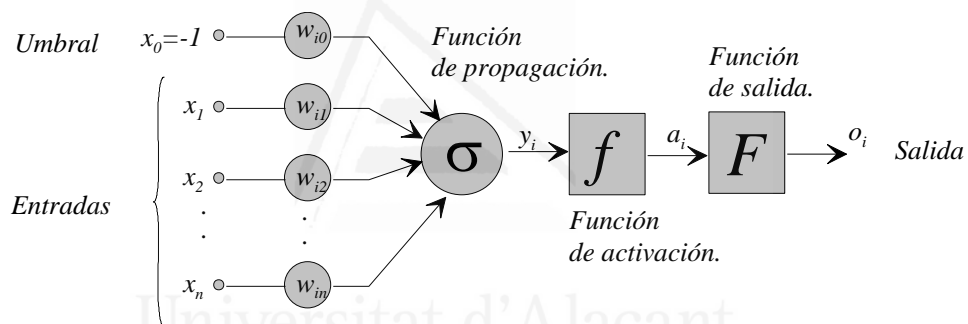


FIGURA VI-6: Modelo general de una neurona artificial.

a) Entradas.

Los valores que llegan del exterior a la neurona pueden ser binarios o continuos dependiendo del modelo y aplicación de la red. Indicaremos mediante x_i la entrada i de la neurona. Generalmente los modelos de neuronas incluyen una entrada adicional de valor constante x_0 con el fin de introducir el umbral de activación de forma sencilla.

a) Conexiones [FIGURA VI-7]:

El peso sináptico representa la intensidad de interacción entre neuronas. Cuanto mayor es su valor mayor es la influencia de la neurona presináptica en la postsináptica. Además las conexiones son direccionales, es decir, sólo propagan información en un solo sentido.

Nos referiremos a una determinada conexión en particular mediante la letra minúscula w y dos subíndices que mostrarán las neuronas ligadas por la sinapsis. Así w_{ij} indica la conexión entre la neurona postsináptica i y la neurona presináptica j .

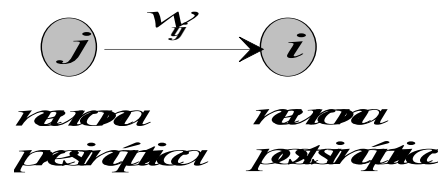


FIGURA VI-7: Representación de la conexión w_{ij} .

b) Regla de propagación $\sigma(w, x(t))$:

Proporciona el valor del potencial postsináptico $y(t)$ de la neurona en función de sus pesos y entradas, $y(t) = \sigma(w, x(t))$. La función más habitual es la suma de los productos de cada entrada por el valor que caracteriza su conexión,

$$y_i = \sum_{j=1}^N w_{ij} x_j(t) - \theta_i$$

donde $j=1,2,3,\dots,N$ total de neuronas postsinápticas.

El término θ_i , denominado umbral de activación, permite que la función de activación no esté centrada en el origen. Una forma elegante de hacerlo es mediante una entrada adicional x_0 de valor constante y de conexión variable w_0 de esta forma:

$$\theta_j = -w_0 x_0$$

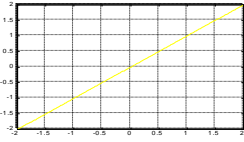
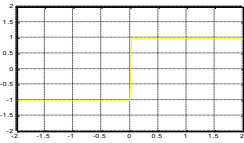
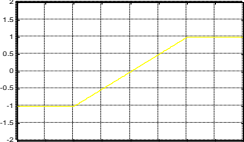
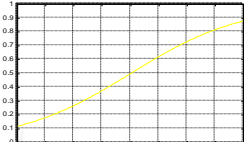
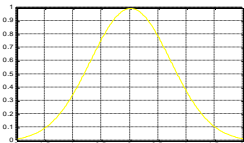
y por tanto el potencial postsináptico se puede escribir de una forma compacta

$$y_i(t) = \sum_{j=0}^N w_{ij} x_j(t)$$

c) Función de activación $f(a(t-1), y(t))$:

Proporciona el estado de activación actual $a(t)$ de la neurona, en función de su estado anterior y de su potencial postsináptico actual, $a(t) = f(a(t-1), y(t))$. Para la mayoría de modelos se considera que el estado actual de la neurona no depende de su estado anterior sino únicamente del actual $a(t) = f(y(t))$. La función $f(\cdot)$ se suele considerar determinista y en la mayor parte de los modelos es monótona creciente y continua. En general los algoritmos de aprendizaje requieren que la función de activación cumpla la condición de ser derivable. En la [TABLA VI-1] se dan las funciones de activación más utilizadas en aplicaciones prácticas.

TABLA VI-1: Tabla de las funciones de activación más comunes.

	<i>Función</i>	<i>Representación</i>
Identidad	$f(x) = x$	
Escalón	$f(x) = \begin{cases} +1 & \text{si } x \geq 0 \\ -1 & \text{si } x < 0 \end{cases}$	
Lineal a tramos	$f(x) = \begin{cases} 1 & \text{si } x \geq 1 \\ x & \text{si } -1 \leq x < 1 \\ -1 & \text{si } x < -1 \end{cases}$	
Sigmoidea	$f(x) = \frac{1}{1 + e^{-x}}$	
Gausiana	$f(x) = e^{-x^2}$	

e) Función de salida F(a(t)):

Proporciona la salida actual o(t) de la neurona en función de su estado de activación, o(t)= F(a(t)). Muy frecuentemente la función de salida es simplemente la identidad, F(x)=x, de modo que el estado de activación de la neurona se considera como la propia salida o(t)= f(y(t)).

A continuación se dará una clasificación de las redes en función del número de capas que posee y la interconexión de los nodos o, dicho de otra manera, en función de su arquitectura o topología.

D.2.- Arquitectura de redes neuronales:

Se denomina arquitectura a la organización y disposición de neuronas formando agrupaciones o capas que comparten características. La estructura de la red determina su comportamiento y está muy relacionada con el algoritmo utilizado para su entrenamiento.

A la hora de clasificar las redes en función de su topología se suele distinguir entre redes de una capa también llamadas monocapa y las que tienen más de una o multicapa.

a) Redes monocapa [FIGURA VI-8]:

En este tipo de redes solamente tenemos una capa de nodos y por tanto las conexiones son laterales. También pueden existir conexiones de una neurona consigo misma (autorecurrentes).

Típicamente estas redes se utilizan en tareas relacionadas con lo que se conoce como autoasociación, es decir, la red responde con el dato almacenado más parecido al de entrada. Generalmente se utilizan para filtrar o reconstruir las informaciones de entradas distorsionadas y para problemas de optimización. Una de las redes monocapas más conocida es la red de Hopfield²⁸⁴.

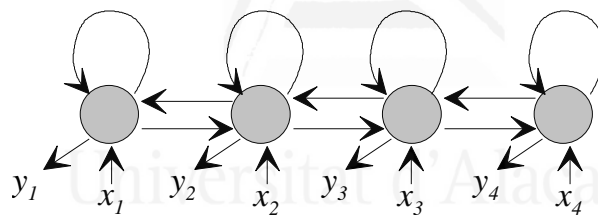


FIGURA VI-8: Ejemplo de red monocapa.

b) Redes multicapa:

En las redes multicapa se distinguen tres tipos de capas; de entrada, de salida y ocultas. Una capa de entrada está compuesta por neuronas que reciben datos o señales procedentes del entorno. Una capa de salida es aquella cuyas neuronas proporcionan la respuesta de la red neuronal y una capa oculta es aquella que no tiene una conexión directa con el entorno. La función de la capa oculta es intervenir entre la entrada y la salida de la red. Añadiendo capas ocultas la red es capaz de extraer estadísticas de alto orden de tal forma que la red adquiere una perspectiva global a pesar de su conectividad local gracias a un conjunto extra de conexiones y la dimensionalidad extra de las interacciones neuronales proporcionada por las capas ocultas.

En las redes multicapa se distingue entre las redes que sólo tienen conexiones hacia delante y las que poseen, además, conexiones hacia atrás.

b.1) RNA con conexiones hacia delante (feedforward) [FIGURA VI-9].

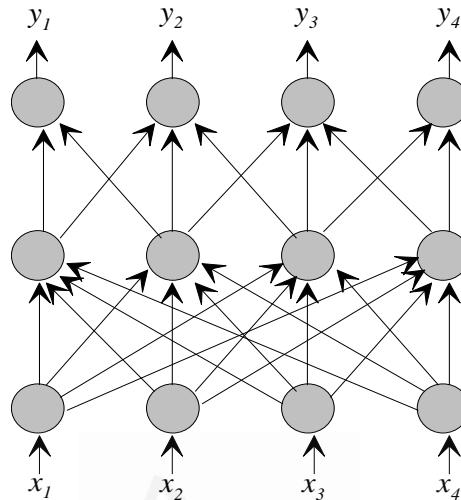


FIGURA VI-9: Red multicapa con conexiones hacia delante.

Generalmente todas las neuronas de una capa reciben señales de entrada de otra capa anterior, más cercana a la entrada de la red, y envían las señales de salida a una capa posterior, más cercana a la salida de la red. Todas ellas son especialmente útiles en tareas de reconocimiento, clasificación de patrones y como aproximador de funciones. Las redes multicapa más conocidas son el perceptrón multicapa y el SOM (*Self Organizing Map*).

b.2) RNA con conexiones hacia atrás -retroconversiones- o recurrentes (feedback) [FIGURA VI-10]

Las redes recurrentes se caracterizan en que al menos tienen un lazo de realimentación. En este tipo de redes circula información tanto hacia delante como hacia detrás durante el funcionamiento de la red. Las conexiones hacia detrás tienen un profundo impacto en la capacidad de aprendizaje de las redes y en su rendimiento. Más aún, los lazos de realimentación implican un comportamiento dinámico no lineal en virtud de las no linealidades de las neuronas. Esto juega un papel fundamental en el almacenamiento de información de este tipo de redes.

Algunas de las redes de este tipo tienen un funcionamiento basado en lo que se conoce como resonancia, de tal forma que las informaciones en la primera y segunda capa interactúan entre sí hasta

alcanzar un estado estable. Ejemplos de redes recurrentes son ART y el Neocognitrón.

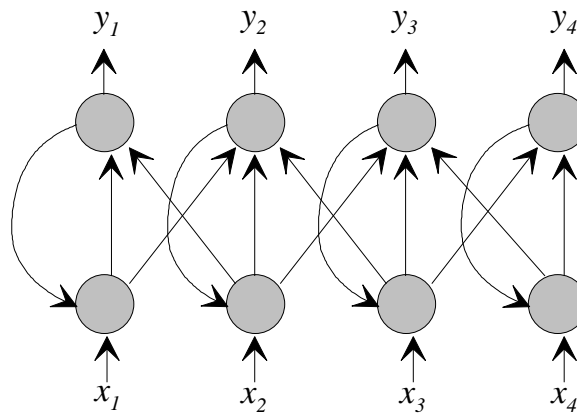


FIGURA VI-10: Red con conexiones hacia atrás.

E. - Características:

Debido a su constitución y fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro²⁵⁴.

E.1.- No linealidad:

Una neurona es un elemento no lineal por lo que una interconexión de ellas (RNA) también será un dispositivo no lineal. La no linealidad es una importante propiedad, particularmente si el responsable de la generación de los datos de entrada es no lineal.

E.2.- Capacidad de establecer relaciones entrada-salida.

Una de las ramas principales de investigación en el campo de las redes neuronales es lo que se conoce como aprendizaje supervisado. La red en su proceso de aprendizaje establece la relación entre las entradas y las salidas de tal manera que, ante entradas desconocidas, es capaz de dar una respuesta "aproximada".

Existen dos formas primarias de realizar esta asociación entre entrada y salida que se corresponden con la naturaleza de la información almacenada en la red. Una sería la denominada heteroasociación que se refiere al caso en el que se relacionan parejas de datos, de tal forma que cuando se presente cierta información de entrada deberá responder generando la correspondiente salida asociada. Otra se conoce como autoasociación donde la red aprende ciertas informaciones de tal manera que cuando se le presenta una información de entrada realizará una autocorrelación respondiendo con uno de los datos almacenados, el más parecido al de entrada.

E.3.- Adaptatividad.

La capacidad de aprendizaje adaptativo es una de las características más atractivas de las redes neuronales. Aprenden a llevar a cabo ciertas tareas mediante un entrenamiento con ejemplos ilustrativos. No es necesario que elaboremos modelos a priori ni necesitamos especificar funciones de distribución de probabilidad.

E.4.- Tolerancia a fallos.

Las redes neuronales típicamente son sistemas de computación robustos o tolerantes a fallos. Esto es posible gracias a que las redes neuronales son sistemas que almacenan la información aprendida de forma distribuida en las conexiones entre neuronas de esta forma se permite el fallo de algunos elementos individuales (neuronas) sin alterar significativamente la respuesta del sistema total.

E.5.- Posibilidad de implementación en VLSI.

Una de las prioridades para la mayoría de las áreas de aplicación es la necesidad de realizar procesos con gran cantidad de datos de forma muy rápida. Las redes neuronales se adaptan bien a estos trabajos pues procesan en paralelo y tienen la posibilidad de implementación en silicio. Esta disposición permite que estos sistemas puedan ser aplicados como sistemas de cómputo en tiempo real.

E.6.- Uniformidad de análisis y diseño.

En todos los dominios de aplicación de las redes neuronales se usa la misma notación, además todas las redes neuronales tienen como elementos básicos las neuronas por lo que es posible enunciar teorías conjuntas para los diferentes algoritmos y aplicaciones de las redes neuronales.

F.- Aprendizaje:

La propiedad más interesante de las redes neuronales es la capacidad para aprender de su entorno y mejorar su comportamiento a través del aprendizaje. Aprendizaje, en este contexto, está definido como un **cambio en los valores de las conexiones** que resultan de la captura de información que, posteriormente, puede ser recuperada[190].

La red aprende de su entorno a través de un proceso iterativo de ajuste de los pesos. De una manera general la expresión para la actualización de los pesos será de la forma:

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}$$

donde w_{ij} muestra el valor de la conexión entre la neurona i y la neurona j y n indica el momento de la actualización.

Hay varios tipos de algoritmos de aprendizaje. Las variaciones entre ellos radican principalmente en la manera de calcular la variación de los pesos $\Delta w_{ij}(n)$.

Una primera distinción entre los posibles procesos de aprendizaje estriba en la disponibilidad o no de un agente externo que supervise el entrenamiento. Este hecho establece una primera clasificación en los métodos de aprendizaje:

F.1.- Aprendizaje Supervisado.

En el proceso de aprendizaje supervisado [FIGURA VI-11] se dispone de un maestro que dirige el ajuste de pesos. Existen tres modalidades en función de la información exterior que maneja el maestro: aprendizaje por error, por refuerzo y aprendizaje estocástico.

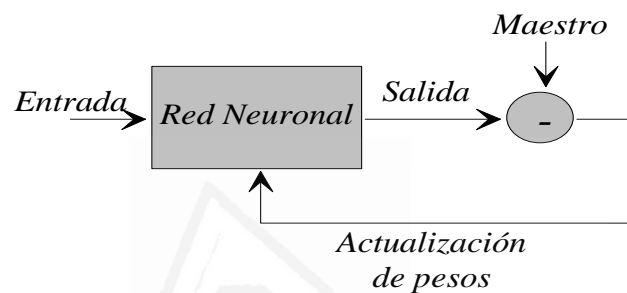


FIGURA VI-11: Esquema del aprendizaje supervisado.

a) Aprendizaje por corrección de error:

El algoritmo por corrección de error consiste en ajustar los pesos de las conexiones en función de la diferencia entre los valores deseados y los obtenidos en la salida de la red. El algoritmo de entrenamiento más famoso por corrección de error, el *backpropagation*, está fundamentado en la optimización de una función coste que representa el error cometido por la red. La actualización de los pesos se lleva a cabo a partir de información proporcionada por el gradiente de la función coste. En este caso la expresión de actualización de los pesos es

$$w_{ij}(n+1) = w_{ij}(n) + \eta \delta_i o_j$$

donde η es una constante que determina el ritmo de aprendizaje, x_j es la entrada j de la neurona, y δ_i es la proporción del error cometido por la red que es asignado a ese nodo i por el algoritmo *backpropagation*.

b) Aprendizaje por refuerzo:

En este caso durante el entrenamiento no se indica exactamente la salida que se desea que proporcione la red ante una determinada entrada; la función del supervisor se reduce a indicar mediante una señal de refuerzo si la salida obtenida en la red se ajusta a la deseada y en función de ello se ajustan los pesos. La ecuación general del aprendizaje por refuerzo es:

$$w_{ij}(n+1) = w_{ij}(n) - \alpha \cdot r(n) \cdot e_{ij}(n)$$

donde α es el factor de aprendizaje, r es un escalar que indica si ha habido acierto o fallo y e_{ij} es la elegibilidad (*elegibility*) del peso w_{ij} . Depende de la distribución de probabilidad que es usada para determinar si el valor de salida de la neurona es igual al valor deseado. La elegibilidad es una especie de memoria; es grande si la señal de entrada a la neurona x_j y la salida o_i están relacionadas en el tiempo.

Mientras que el sistema no falla, no existe refuerzo ($r=0$) y por tanto no se produce ajuste de pesos. Con tal de evitar este inconveniente se introduce un sistema de salida continua, que aprende a dar la predicción de una futura penalización; de esta forma se consigue tener siempre una señal de refuerzo y el aprendizaje se mejora notablemente. Este tipo de aprendizaje, por sus características, es especialmente adecuado para tareas de control ya que en estas tareas generalmente se dispone de muy poca información del sistema con el que se trata.

c) Aprendizaje estocástico:

Consiste básicamente en realizar cambios aleatorios en los valores de los pesos de las conexiones de la red y evaluar su efecto a partir del objetivo deseado y distribuciones de probabilidad. Se suele hacer una analogía con términos termodinámicos y el aprendizaje consistiría en realizar cambios aleatorios y determinar la energía del sistema; si ésta disminuye se acepta el cambio, si no, el cambio se aceptaría en función de una preestablecida distribución de probabilidades.

F.2.- Aprendizaje no supervisado.

No requieren la influencia externa [FIGURA VI-12] para ajustar los pesos de las conexiones entre sus neuronas. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta. Por ello suele decirse que estas redes tienen la capacidad de auto-organizarse. Deben encontrar las características, regularidades, correlaciones o categorías que se puedan establecer entre los datos que se presenten en su entrada.

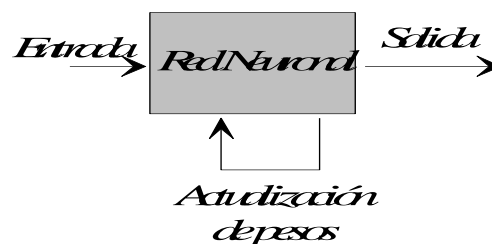


FIGURA VI-12: Esquema del aprendizaje no supervisado.

Dentro del aprendizaje no supervisado se distingue entre dos tipos.

a) Aprendizaje Hebbiano:

El aprendizaje hebbiano consiste básicamente en el ajuste de los pesos de las conexiones de acuerdo con la correlación entre los valores de entrada y salida de cada neurona. Existen diferentes tipos en función de las limitaciones de los pesos, o particularidades que deseamos obtener de la secuencia de entrada a la red.

a.1) Aprendizaje Hebbiano Original:

Coincide con el postulado formulado por D.O. Hebb en 1949²⁴⁸. La actualización de los pesos se corresponde con la correlación entre los valores de entrada y la salida de cada neurona.

$$w_{ij}(n+1) = w_{ij}(n) + o_i x_j$$

De esta forma cuando la entrada x_j y la salida o_i se comportan de forma similar la conexión se refuerza, penalizándose en otro caso.

a.2) Aprendizaje de componentes principales.

El análisis de las componentes principales es un método que permite la reducción de las dimensiones del espacio de variables en el que se trabaja. El primer aprendizaje de componentes principales fue realizado por Oja (1982) el cual razonó que el aprendizaje hebbiano con algunas modificaciones servía para extraer las componentes principales de los datos de entrada.

La ecuación propuesta para la actualización de las conexiones es²⁸⁵

$$w_{ij}(n+1) = w_{ij}(n) + o_j (\alpha x_i - \beta o_j w_{ij})$$

donde α y β son dos constantes de proporcionalidad.

Se trata una topología de red que utiliza el aprendizaje por corrección de error (backpropagation) y a la vez intercala capas donde se determinan las componentes principales reduciendo así, y de manera progresiva, el número de nodos necesarios en las capas ocultas²⁸⁵.

b) Aprendizaje competitivo.

En el aprendizaje competitivo se crean clases de los patrones de entrada automáticamente. En las redes con este tipo de aprendizaje suele decirse que las neuronas compiten unas con otras con el fin de llevar a cabo una tarea dada.

En este tipo de redes se pretende que al presentar cierta información de entrada sólo una de las neuronas de la salida, o un grupo de neuronas, se

active. Para conseguirlo existen conexiones autorrecurrentes de refuerzo y de tipo inhibitoras por parte de las neuronas vecinas. La interacción entre las neuronas de salida se define mediante la función de vecindad. Una posibilidad es la que se muestra en la [FIGURA VI-13] con forma de sombrero.

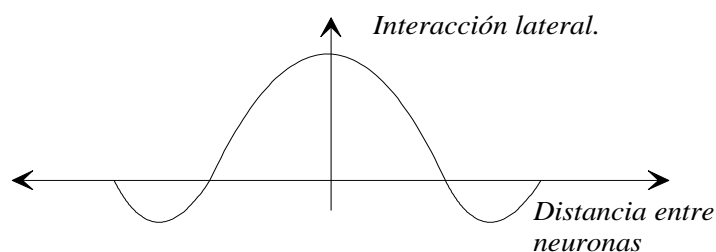


FIGURA VI-13: Interacción entre neuronas de la capa de salida.

Las neuronas ganadoras son las únicas que ajustan sus pesos con tal de modificar el prototipo que representan para incluir a la nueva entrada.

$$w_{ij}(n+1) = w_{ij}(n) + \alpha(n)(x_j - w_{ij})$$

donde $\alpha(n)$ es una función decreciente del tiempo, sin llegar a ser cero. El resultado de esta operación es el movimiento del vector prototipo de la clase hacia el vector de entrada. La red más extendida que emplea este tipo de aprendizaje es el SOM de Kohonen²⁵².

G.- Elección de la RNA más adecuada:

El objetivo, en la presente tesis de licenciatura, es el desarrollo de una RNA que, a partir de las características recogidas por la Puntuación PRISM de las primeras 24 horas, sea capaz de predecir la mortalidad del paciente ingresado en UCIP. Como se ve, ello se traduce a una tarea de clasificación bien definida, y las redes neuronales son una opción excelente para su utilización en esta aplicación, pues no parten de restricciones respecto de los datos de partida, ni imponen supuestos (ej. distribución gaussiana) y, si se implementan en el vehículo adecuado - como a través de Internet- son más fáciles de emplear.

Pero, ¿qué tipo de red es adecuada para solucionar el problema del presente trabajo? Se ha visto que las redes neuronales artificiales pueden clasificarse según su arquitectura, o bien, según el método de aprendizaje. La misión de la red consistirá en determinar, lo más fielmente posible, el grupo al que pertenece un niño según su puntuación en el score PRISM. [FIGURA VI-14]. Dicho de otra forma, lo que se busca es encontrar la relación que existe entre la variable puntuación PRISM y las clases representativas de su evolución (Vivo / Muerto). Para obtener una correspondencia de este tipo necesitamos al menos dos capas. La

primera contendría las variables dependientes y la segunda los grupos de clasificación. Esto solamente puede suceder cuando tenemos redes multicapa.

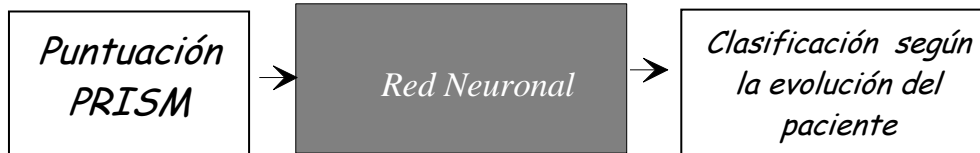


FIGURA VI-14: Esquema de la tarea a realizar por la red neuronal en el problema del PRISM.

El proceso de entrenamiento de la red estará fundamentado en la presentación de los casos de niños ingresados en la UCIP del Hospital Central de Asturias hasta que la red aprenda su cometido. El modo de aprendizaje será de tipo supervisado porque se tiene la posibilidad de disponer de una medida exacta del error cometido por la red. Esto es así, dado que se sabe de antemano el resultado que debe dar la red cuando presenta un nuevo caso (en fase de entrenamiento), es decir, a partir de la respuesta de la red se puede determinar cuánto se ha equivocado y actuar en consecuencia con el fin de mejorar el comportamiento de la red.

En la segunda fase del trabajo, se evaluará el rendimiento diagnóstico de la RNA en una cohorte de validación diferente: la de los niños ingresados en la UCIP del Hospital Infantil "La Fe" de València. Como se comenta en los métodos, la evaluación se hace con la metodología clásica, y se hace una propuesta de evaluación con metodología bayesiana.

ANEXO VII: USO CLÍNICO DE UN TEST DIAGNÓSTICO: MODELO DE PAUKER-KASSIRER MODIFICADO POR LATOUR

El uso adecuado de una prueba diagnóstica se basa en la asunción de un axioma: desde el punto de vista clínico una prueba diagnóstica es útil sólo si nos induce a tomar las decisiones terapéuticas adecuadas en un ambiente de incertidumbre. Debemos pedir una prueba diagnóstica sólo si es capaz de **hacernos cambiar la decisión terapéutica** que habíamos tomado antes de pedirla.

Pauker y Kassirer, en una serie ya antigua de artículos^{286,287,288}, desarrollaron un abordaje sistemático muy sencillo para ayudar a los clínicos a un uso correcto de las pruebas diagnósticas a pie de la cama del enfermo²⁸⁹. Está basado en la Teoría Matemática de la Decisión y en el Cálculo de Probabilidades. Nuestra propuesta para el uso clínico correcto de las pruebas diagnósticas, se basa en la modificación que el Profesor Jaime Latour ha hecho²⁹⁰ del modelo Pauker-Kassirer de análisis de decisiones clínicas.

Lo explicaremos con un ejemplo concreto, adaptado a un escenario determinado, que se repite ininidad de veces en la Puerta de Urgencias de un hospital pediátrico de nuestro país. El escenario es el siguiente:

ESCENARIO: Llega a PU un niño de 5 ½ meses de edad. Su madre lo trae porque, estando previamente bien, desde hace 8 horas, tiene fiebre de 39'3 °C que cede bien con antitérmicos, aunque reaparece. Tras bajarle la temperatura, exploras al niño, que está despierto y contento, sonrosado (SaO₂=97%) y bien hidratado. No encuentras signos de focalidad infecciosa ni disnea ni estertores respiratorios, la orofaringe es normal y no tiene signos meníngeos, exantemas ni petequias. Le practicas un Hemograma (Leucocitos 8.900 /mm³ con fórmula normal), un Urinoanálisis (Nitritos negativos, resto normal), y una PCR que es de 143 mg/Lit.

La pauta de tu hospital te recomienda extraer hemocultivo y, bajo la vigilancia estricta de los padres (que te parecen fiables), administrar una dosis IM/IV de 50 mg/kg de Ceftriaxona, pero advirtiéndole a los familiares que si empeora vuelvan para re-evaluarlo, y que si sigue todo bien, lo lleven mañana a su pediatra.

Un compañero ha hecho un rotatorio por otro hospital, y te cuenta que allí, en estos casos practican un Test de Procalcitonina (PCT) para confirmar la sospecha de bacteriemia oculta, antes de proceder a administrar antibioticoterapia (que podría producirle al niño alteraciones gastrointestinales innecesarias).

El clínico, enfrentado a este paciente particular, decide realizar una búsqueda bibliográfica rápida en el ordenador del Servicio de urgencias, y recupera el siguiente artículo sobre el Test de Procalcitonina:

Annick Galletto-Lacour, Samuel A. Zamora, Alain Gervais: Bedside Procalcitonin and C-Reactive Protein Test in Children With Fever Without Localizing Signs of Infection Seen in a Referral Center. *Pediatrics* 2003; 112: 1054-60.

El pediatra hace una lectura crítica del mismo, y comprueba su gran validez. Se trata de un estudio prospectivo realizado en pacientes europeos aproximadamente del mismo espectro clínico de la situación que plantea el escenario, con niños totalmente intercambiables con el que tiene delante, en el que existe una comparación ciega e independiente del Test de PCT con un "patrón oro" (una definición operativa de enfermedad bacteriana grave que se aplica a todos los pacientes del estudio al final de la evolución de su proceso patológico), y en el que la aplicación, concordancia y reproducibilidad del test está perfectamente especificada. El laboratorio del hospital dispone de la misma técnica que se describe en el artículo y es posible pedirla de urgencia.

El resultado principal del trabajo está reproducido en la tabla siguiente, en la que se especifican las Razones de Verosimilitud de la PCT (categorizada en negativa [< 0.5 ng/mL], dudosa [$0.5-2$ mg/mL], y positiva [> 2 mg/mL]), la PCR y el recuento de Leucocitos del Hemograma, con sus Intervalos de Confianza al 95%:

TABLE 3. LRs for Selected Range of Values of PCT, CRP, and Leukocyte Counts and Posttest Probability of SBI in Children With FWS

	<i>n</i>	LR (95% CI)	Posttest Probability (%)
PCT			
<0.5 ng/mL	54	0.09 (0.02–0.36)	3
0.5–2	26	2.8 (1.49–5.33)	54
>2	19	5.2 (2.20–12.42)	68
CRP			
<40 mg/L	61	0.26 (0.13–0.54)	10
40–100	22	2.0 (1.04–4.01)	45
>100	16	14.5 (3.46–60.70)	86
Leukocyte			
<15 G/L	66	0.65 (0.44–0.97)	21
15–20	15	1.6 (0.63–4.11)	40
>20	18	2.4 (1.07–5.46)	49

Pretest probability: 29%.

Ante la situación del escenario, y con la información que ha obtenido de la literatura, el clínico se plantea básicamente dos preguntas que van a dirigir su actitud:

1. ¿Si la Procalcitonina es negativa (< 0.5 ng/ml) se puede descartar en este momento una evolución posterior a sepsis-meningitis en este niño?
2. ¿Pautarías Ceftriaxona IV a este niño?

Vamos a aplicar a esta situación clínica particular el modelo Pauker-Kassirer-Latour. Básicamente, consiste en realizar tres pasos sucesivos: primero determinar el umbral de acción (UA); segundo, determinar -en las condiciones estrictas del paciente del escenario- cuál es la probabilidad de enfermedad antes de hacer la prueba y, por último, calcular la probabilidad posprueba de enfermedad

dados los resultados del test. Si el resultado del test es capaz de llevar la probabilidad de enfermedad más allá del valor umbral, el uso del test está clínicamente justificado (en las condiciones del escenario) porque va a ser capaz de cambiar nuestra actitud terapéutica.

A.- Primero: Determinar el Umbral de Acción (UA):

Lo primero que debemos hacer es determinar el UA para la situación del escenario. Esto es: en un niño con síndrome febril sin foco como este que tenemos delante, que la inmensa mayoría de veces se debe a una viriasis banal pero que puede tratarse de una bacteriemia oculta, ¿A partir de que probabilidad de evolucionar a sepsis-meningitis deberíamos tomar una actitud terapéutica y pautar Ceftriaxona IV?

Para el planteamiento Pauker-Kassirer-Latour, el UA es un cociente Riesgo/Beneficio: el Riesgo esperado de aplicar el tratamiento y el Beneficio que esperamos obtener con él. Obviamente, por razones éticas, siempre Riesgo < Beneficio. Por ello el valor del UA oscila entre 0 y 1, y posee las propiedades matemáticas de una probabilidad:

$$UA = \frac{\text{Riesgo}}{\text{Beneficio}} = \frac{\text{Daño esperado con el tratamiento}}{\text{Beneficio esperado con el tratamiento}}$$

Siendo:

- Riesgo = Probabilidad de que el tratamiento produzca efectos adversos (EA) multiplicada por el impacto que esos EAs producen en el enfermo, medido en una escala de 0 a 1 = p(EA)*impacto(EA)
- Beneficio = Probabilidad de que el tratamiento cure la enfermedad multiplicada por el impacto, medido en la misma escala en la que se ha evaluado el impacto de los EAs, que una curación tiene para el paciente = p(cura)*impacto (cura)

Para poder calcular el UA, debemos consultar la literatura. Se sabe²⁹¹ que en niños de entre 3 y 6 meses que se presentan en urgencias con fiebre sin foco (FSF), el riesgo de bacteriemia oculta es del 4%, y de los que tienen bacteriemia oculta, el 30% evolucionan a sepsis-meningitis. Por ello puede calcularse que el riesgo global de evolución a sepsis-meningitis ante un niño como el del escenario es:

$$p(\text{sep-men/FSF}) = 0.04 * 0.3 = 0.012$$

Un ensayo aleatorizado y controlado con placebo²⁹² demostró que la Ceftriaxona parenteral, en niños mayores de 3 meses con bacteriemia oculta, es capaz de reducir la evolución a sepsis-meningitis a sólo un 8.3%. Esto es:

$$p(\text{sep-men/FSF+ceftriaxona}) = 0.04 * 0.083 = 0.0033$$

Con este dato podemos calcular el tamaño del efecto (RAR, Reducción absoluta del Riesgo) de la ceftriaxona para evitar la evolución a sepsis meningitis en los pacientes con FSF como el nuestro. Nuestra $p(\text{cura})$ será exactamente esta RAR:

$$p(\text{cura}) = \text{RAR} = p(\text{sep-men}/\text{FSF}) - p(\text{sep-men}/\text{FSF} + \text{ceftriaxona}) = \\ = 0.012 - 0.0033 = 0.0087 \quad \text{es decir, NNT} = 115 \text{ pacientes}$$

Y ahora estimamos el impacto que tiene para el paciente del escenario el que se evite para él la evolución a sepsis-meningitis. Podemos asumir que

$$\text{Impacto evitar evolución a sepsis-meningitis} = 1 = \text{impacto}(\text{cura})$$

En el mismo ensayo clínico²⁹² podemos observar que el principal efecto adverso del tratamiento con ceftriaxona son las alteraciones gastrointestinales (diarrea, vómitos...), que se producen en el 21.2% de los niños a los que se les administra. Pero que se ve también que el 11.7% de los pacientes que reciben placebo presentan esas mismas alteraciones gastrointestinales (que probablemente eran la causa de su síndrome febril). Así que se puede calcular que la potencia que tiene el fármaco para producir los EAs es el tamaño de ese efecto (AAR, Aumento absoluto del riesgo):

$$p(\text{EA}) = \text{AAR} = p(\text{diarrea}/\text{FSF} + \text{ceftriaxona}) - p(\text{diarrea}/\text{FSF}) = \\ = 0.212 - 0.117 = 0.095 \quad \text{es decir, NNH} = 11 \text{ pacientes}$$

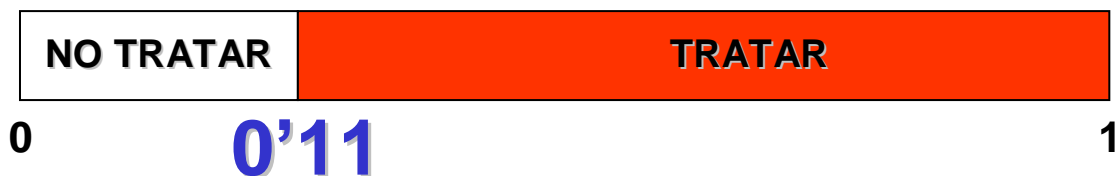
Y ahora, si 1 es el valor que tiene para un paciente evitar su evolución a sepsis-meningitis, ¿cuál es el impacto de producirle alteraciones gastrointestinales? Podemos asumir que será una centésima parte:

$$\text{Impacto producir diarrea, vómitos} = 0.01 = \text{impacto}(\text{EA})$$

Ya estamos pues en disposición de calcular el UA para nuestro escenario:

$$UA = \frac{\text{Riesgo}}{\text{Beneficio}} = \frac{p(\text{EA}) \times \text{impacto}(\text{EA})}{p(\text{cura}) \times \text{impacto}(\text{cura})} = \frac{0.095 \times 0.01}{0.0087 \times 1} = 0.11$$

Vamos a representar este US en un diagrama de barra en el que se localiza como una probabilidad sobre una escala de 0 al 1:



El diagrama representa que nosotros, ante un paciente del escenario, hemos decidido que vamos a aplicar un tratamiento con Ceftriaxona IV cuando estemos convencidos de que la probabilidad de que evolucione a sepsis-meningitis sea superior al 11%. Mirando el gráfico, recordaremos que sólo está justificado utilizar una prueba diagnóstica cuando esta sea capaz de cambiar nuestra actitud terapéutica. Es decir, si estamos decididos a tratar al paciente (pensamos que su probabilidad es mayor de 0.11), y el resultado de la prueba puede hacer que su probabilidad baje por debajo de ese umbral y haga que no lo tratemos, y viceversa.

B.- Segundo: Estimar, para el paciente del escenario, la probabilidad de enfermedad antes de realizar la prueba:

Una vez fijado cuál va a ser nuestro umbral de acción, lo siguiente que debemos hacer es determinar, cuál es la probabilidad pre-prueba de que el **paciente del escenario** presente la enfermedad. En nuestro caso particular, la probabilidad de que nuestro niño de 5 meses y medio con fiebre sin foco pero PCR = 143 mg/L, evolucione a desarrollar una sepsis-meningitis.

El paciente de nuestro escenario es un lactante de entre 3 y 6 meses con FSF: Tiene fiebre de 39.3°C rectal, pero no presenta aspecto tóxico, no se aprecia focalidad alguna en la exploración física, y tiene un recuento leucocitario normal (Leucos 8900 /mm³). Hemos encontrado en la literatura que

$$p(\text{sep-men/FSF}) = 0.04 * 0.3 = 0.012; \quad \text{es decir [odds} = p/(1-p)]$$

$$\text{Odds}(\text{sep-men/FSF}) = 0.012/0.988 = 0.012 = \text{Odds Pre-PCR}$$

Pero sabemos además que el niño de nuestro escenario tiene una PCR = 143 mg/L. En la tabla 3 del artículo que hemos analizado (ver más arriba), hemos encontrado que la RV para una PCR > 100 mg/L es: $RV(\text{PCR} > 100) = 14.5$. Con ello podemos estimar la probabilidad de evolucionar a sepsis-meningitis, en el **paciente de nuestro escenario**, aplicando el TEOREMA de BAYES:

$$\text{Odds}(\text{sep-men/escenario}) = \text{Odds Post-PCR} = \text{Odds Pre-PCR} * RV(\text{PCR} > 100) =$$

$$= 0.012 * 14.5 = 0.176 ; \quad \text{es decir [p} = \text{odds}/(1+\text{odds})]$$

$$p(\text{sep-men/escenario}) = 0.176 / 1.176 = 0.15$$

La probabilidad de que nuestro lactante evolucione a sepsis-meningitis, estimada antes de que le practiquemos la prueba de procalcitonina es del 15%, claramente superior a nuestro UA.

Prob pre-PCT = 0'15



B.- Tercero: Calcular, la probabilidad de enfermedad después de realizar la prueba al paciente del escenario:

Finalmente hemos de calcular, utilizando los índices de rendimiento diagnóstico de nuestro estudio, cual será la probabilidad de que el **paciente de nuestro escenario** presente la enfermedad a la luz de los posibles resultados de la prueba diagnóstica. Recordemos que la prueba será útil sólo si es capaz de disminuir la probabilidad de evolución a sepsis-meningitis hasta una cifra inferior al 11%. Sólo en ese caso tendrá valor clínico, porque hará que **NO TRATEMOS** con ceftriaxona a nuestro lactante. En caso contrario, si es incapaz de disminuir la probabilidad a un valor inferior al 11% (si, sea cual sea el resultado de la prueba de procalcitonina, el **paciente de nuestro escenario** siempre va a tener una probabilidad de sepsis-meningitis mayor del 11%) el solicitar la prueba no tiene ningún valor para nuestra situación clínica particular.

Aplicando de nuevo el TEOREMA de BAYES con los datos de la tabla 3 de nuestro estudio, si el resultado de PCT fuera positivo ($PCT > 2 \text{ ng/mL}$) o dudoso ($0.5 \text{ ng/mL} < PCT < 2 \text{ ng/mL}$), la probabilidad de evolución a sepsis-meningitis en el lactante de nuestro escenario sería superior al UA, y por tanto nuestra actitud terapéutica no cambiaría:

- PCT positiva: $RV(PCT \text{ positiva}) = RV(PCT > 2 \text{ ng/mL}) = 5.2$ luego
 $Odds(\text{sep-men/escenario}) * RV(PCT \text{ positiva}) = Odds(\text{sep-men/Post-PCT})$
 $Odds(\text{sep-men/Post-PCT}) = 0.176 * 5.2 = 0.9152$;
 es decir $[p = odds/(1+odds)]$:
 $p(\text{sep-men/post-PCT}) = 0.91/1.91 = 0.48 (> 0.11: \underline{\text{TRATAR}})$
- PCT dudosa: $RV(PCT \text{ dudosa}) = RV(0.5 \text{ ng/mL} < PCT < 2 \text{ ng/mL}) = 2.8$ luego
 $Odds(\text{sep-men/escenario}) * RV(PCT \text{ dudosa}) = Odds(\text{sep-men/Post-PCT})$
 $Odds(\text{sep-men/Post-PCT}) = 0.176 * 2.8 = 0.4928$;
 es decir $[p = odds/(1+odds)]$:
 $p(\text{sep-men/post-PCT}) = 0.49/1.49 = 0.33 (> 0.11: \underline{\text{TRATAR}})$

Sin embargo, en el **paciente del escenario**, un resultado negativo ($PCT < 0.5 \text{ ng/mL}$) de la prueba de procalcitonina **ES CAPAZ DE CAMBIAR NUESTRA ACTITUD TERAPÉUTICA**. Veámoslo:

- PCT negativa: $RV(PCT \text{ negativa}) = RV(PCT < 0.5 \text{ ng/mL}) = 0.09$ luego
 $Odds(\text{sep-men/escenario}) * RV(PCT \text{ negativa}) = Odds(\text{sep-men/Post-PCT})$
 $Odds(\text{sep-men/Post-PCT}) = 0.176 * 0.09 = 0.016$;
 es decir $[p = odds/(1+odds)]$:
 $p(\text{sep-men/post-PCT}) = 0.016/1.016 = 0.016 (\ll 0.11: \underline{\text{NO TRATAR}})$

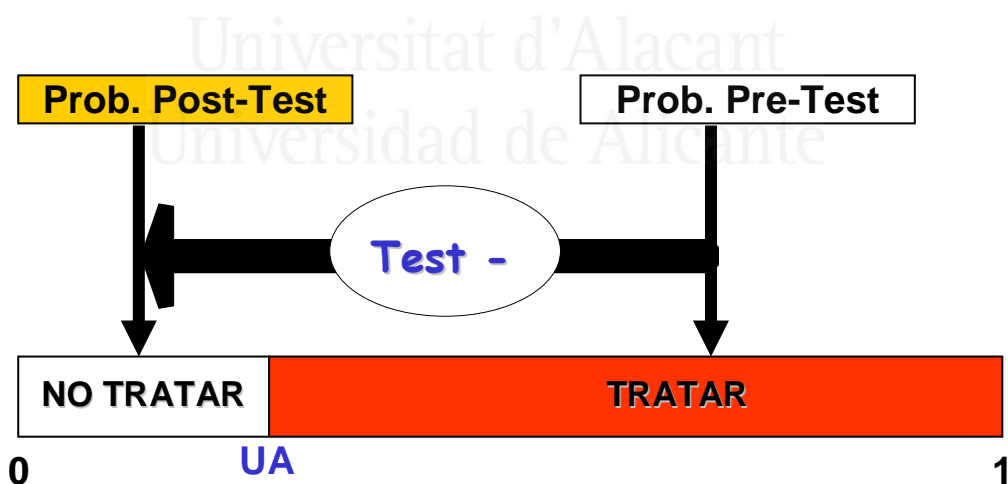
Prob post-PCT = 0'016



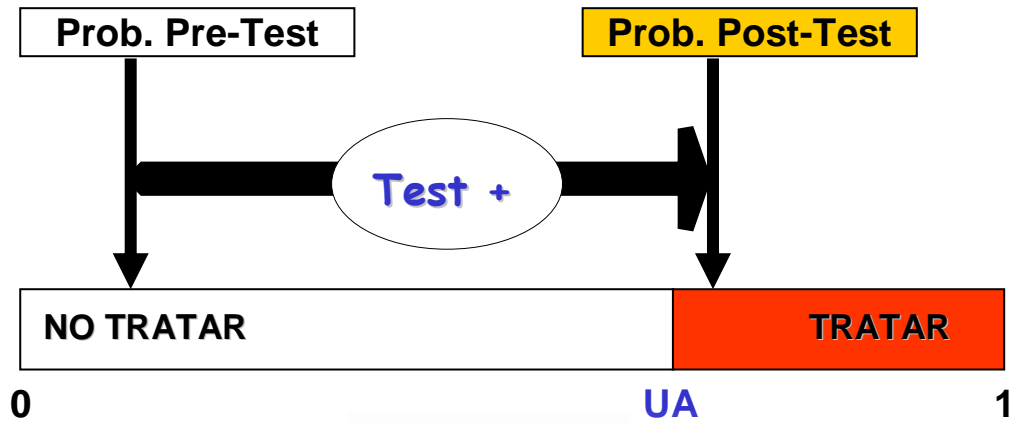
Por ello en la situación de este lactante, tiene sentido clínico que se le solicite una prueba de PCT y que, con su resultado, se decida el tratamiento. Por ello, como respuesta a las preguntas que se había planteado nuestro pediatra de urgencias:

1. Si la Procalcitonina es negativa (< 0.5 ng/ml) se puede descartar en este momento una evolución posterior a sepsis-meningitis en este niño, y
2. Si la Procalcitonina sale negativa, no pautaría Ceftriaxona IV a este niño

Como norma general, en escenarios en los que el UA está muy bajo (lo normal es aplicar tratamiento), suelen ser útiles pruebas diagnósticas cuyos resultados negativos tengan RV(negativo) muy bajas, para que sean capaces de evitar los tratamientos.



En los escenarios contrarios, con UA muy altos en los que lo habitual es no tratar, resultan útiles las pruebas con RV(positivo) muy altos, para que nos induzcan a cambiar de actitud y tratar a los pacientes.



Universitat d'Alacant
Universidad de Alicante



X. - Bibliografía.

Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante

- ¹ Poincaré H. Ciencia e Hipótesis (La Science et l'Hypothèse; 1902). 2ª ed. Madrid: Espasa Calpe, S.A., 2002. pp 214-5.
- ² Feynman RP. Probabilidad e incertidumbre. La visión de la naturaleza a través de la mecánica cuántica (1965). El carácter de la ley física. Tusquets, 2000: 141-163.
- ³ Spodick DH. On experts and expertise: The effect of variability in observer performance. Am J Cardiol 1975; 36: 592.
- ⁴ Alcock JE. The belief engine. Skeptical Inquirer 1995; 19(3):14-18. Disponible On-Line en <http://www.csicop.org/si/9505/belief.html>
- ⁵ Park RL. Ciencia o Vudú: de la ingenuidad al fraude científico. (Woodoo science). 2001 ed. Farigliano (Italia): Grijalbo Mondadori SA, 2001.
- ⁶ Hume D. Compendio de un tratado de la Naturaleza Humana (An abstract of a book lately Publisher entitled A Treatise of Human Nature, 1740). 1ª ed. Ed.: Federico Ruiz Company. Colección cuadernos de Filosofía. Valencia: Ediciones Tilde, 1999.
- ⁷ Hume D. Tratado de la naturaleza humana (A Treatise of Human Nature, 1739-1740). 2ª ed. Fuenlabrada (Madrid): Editorial Tecnos S.A., 1992. p.155.
- ⁸ Hume D. Investigación sobre el conocimiento humano (Enquiry Concerning the Human Understanding, 1748). 11ª ed. Madrid: Alianza Editorial, 1997.
- ⁹ Russell B. Los problemas de la filosofía. 8ª ed. Barcelona: Editorial Labor, 1983.
- ¹⁰ Sala i Martín X. Economía liberal para no economistas y no liberales (Economia liberal per a no economistes i no liberals, 2001). 1ª ed. Barcelona: Plaza&Janés editores S.A., 2002.
- ¹¹ Popper KR. La lógica de la investigación científica (Logik der Forschung, 1934). 10ª ed. Navalcarnero (Madrid): Editorial Tecnos, S.A., 1997.
- ¹² Hempel CG. Filosofía de la ciencia natural (Philosophy of Natural Science, 1966). 17ª ed. Paracuellos de Jarama (Madrid): Alianza Editorial, 1998.
- ¹³ Medawar P. Hipótesis e imaginación (25/Oct/1963). En: Medawar, P. El extraño caso de los ratones moteados y otros ensayos sobre ciencia. Barcelona: Crítica (Grijalbo Mondadori), 1997. pp 31-49.
- ¹⁴ Medawar P. Previsión y predicción (Oct/1982). En: Medawar, P. El extraño caso de los ratones moteados y otros ensayos sobre ciencia. Barcelona: Crítica (Grijalbo Mondadori), 1997. pp 31-49.
- ¹⁵ Von Wright GH. The logical problem of induction. 2ª ed. Oxford. Basil Blackwell & Mott. 1965.
- ¹⁶ de Moivre A. The doctrine of chances; Or a method of calculating the probabilities of events in play (3th ed.1756). 1ª ed. American Mathematical Society, 2000.

¹⁷ Barnard GA. Thomas Bayes' essay towards solving a problem in the doctrine of chances (Reimpresión de Bayes, T. An essay towards solving a problem in the doctrine of chances *The Philosophical Transactions of The Royal Society, 1763; 53: 370-418*). *Biometrika* 1958; 45:293-315.

¹⁸ Kolmogórov AN. Foundations of the theory of probability (Grundbegriffe der Wahrscheinlichkeitrechnung. *Ergebnisse der Mathematik, 1933*). 2ª ed. New York: Chelsea Publishing Co., 1956

¹⁹ Hacking I. An introduction to probability and inductive logic. 1ª ed. New York (U.S.A.): Cambridge University Press, 2001.

²⁰ Stuart Mill JS. *A system of logic*. (1ª edición 1879). 8ª Ed. Londres. Longmans, Green & Co.

²¹ Jeffreys sH. *Scientific Inference* (1st ed 1931). 2ª ed. London, England: Cambridge University Press, 1957.

²² von Mises R. *Probability, Statistics and Truth* (2ª ed. revisada. 1957). 1ª ed. New York: Dover Publications, Inc, 1981

²³ Hacking I. *La domesticación del azar* (The taming of chance). 1ª ed. Barcelona: Editorial Gedisa S.A., 1995.

²⁴ Neyman J. Fiducial arguments and the theory of confidence intervals. *Biometrika* 1941; 32:128-150.

²⁵ Fisher RA. *Statistical Methods for Research Workers*(1st. Ed 1925). 14ª ed. New York: Haffner Publishing Company, 1973.

²⁶ Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika* 1928; 20A:175-240.

²⁷ Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* 1928; 20A:263-294.

²⁸ Neyman J, Pearson ES. On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A* 1933; 231:289-337.

²⁹ Mayo DG. *Error and the growth of experimental knowledge*. 1ª ed. Chicago: The University of Chicago Press, 1996.

³⁰ Giere RN. *Understanding Scientific Reasoning*. 4ª ed. Orlando (Florida): Harcourt Brace College Publishers, 1997.

³¹ Taper ML, Lele SR, (Eds). *The nature of scientific evidence. Statistical, philosophical and empirical considerations*. 1ª ed. Chicago: The University of Chicago Press, 2004.

³² Fisher RA. *The design of experiments* (1st. edition 1935). 8ª ed. New York: Haffner Publishing Company, 1971.

- ³³ Lukasiewicz J. La silogística de Aristóteles desde el punto de vista de la moderna lógica formal. Madrid: Tecnos, 1977.
- ³⁴ Miguel García, F. Popper, el contraste de hipótesis y el método crítico. *Revista Cubana de Salud Pública* 2003; 29(1): 52-60.
- ³⁵ Popper KR. *Conocimiento Objetivo (Objective Knowledge, 1st Ed 1971)*. 4ª ed. Editorial Tecnos (Grupo ANAYA, S.A.), 2001.
- ³⁶ Popper KR. *Conjeturas y Refutaciones (The structure of Science, 1st ed 1961)*. Barcelona: Ediciones Paidós Ibérica S.A., 1991.
- ³⁷ Feynman RP. En busca de nuevas leyes (Messenger Lectures de la Cornell University, 1964). In: *Museu de la Ciència de la Fundació "La Caixa"*, editor. *El carácter de la Ley Física*. Barcelona: Tusquets Editores, 2000: 165-190.
- ³⁸ Diez JA, Moulines CU. Contrastación de hipótesis. En: Diez JA, Moulines CU, ed. *Fundamentos de filosofía de la ciencia*. Barcelona: Editorial Ariel S.A., 1997: 61-90.
- ³⁹ Hill AB. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 1965; 58:295-300.
- ⁴⁰ United States Department of Health, Education and Welfare. Report of the Advisory Committee to the Surgeon General: Smoking and Health. Washington D.C., Public Health Service, 1964.
- ⁴¹ Gordis L, Kleinman J, Kleinman L, et al. Criteria for evaluating evidence regarding the effectiveness of prenatal interventions. En: Merkatz I, Thomson J (eds). *New perspectives on prenatal care*. New York: Elsevier, 1990: 31-38.
- ⁴² Hacking I. *La domesticación del azar (The taming of chance, 1990)*. 1ª ed. Barcelona: Editorial Gedisa S.A., 1995.
- ⁴³ Hacking I. *El surgimiento de la probabilidad (The emergence of probability, 1975)*. 1ª ed. Barcelona: Editorial Gedisa S.A., 1995.
- ⁴⁴ Neyman J. "Inductive behavior" as a basic concept of philosophy of science. *Rev Math Statist Inst* 1957; 25:7-22.
- ⁴⁵ Marshall G. Streptomycin in Tuberculosis Trials Committee of the Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 1948;769-782.
- ⁴⁶ Bernardo JM. Bruno de Finetti en la estadística contemporánea. In: Ríos S, editor. *Historia de la Matemática en el Siglo XX*. Madrid: Real Academia de Ciencias exactas, físicas y naturales, 1998: 63-80.
- ⁴⁷ Bernardo JM. Mètodes estadístics contemporanis en la investigació científica: anàlisi bayesià. *Mètode* 2000; 24:32-34
- ⁴⁸ Schlaifer R. *Probability and Statistics for business decisions*. 1ª ed. New York: McGraw-Hill, 1959.

- ⁴⁹ Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Phil Trans Roy Soc, serie A* 1937; 236:338-380.
- ⁵⁰ Jeffreys sH. *Theory of Probability* (3th. Ed, 1961). Oxford (England): Oxford University Press, 1998.
- ⁵¹ Borel E. A propós d'un traité de probabilité. *Rev Phil* 1924; 98:321-336.
- ⁵² Gillies D. *Philosophical theories of probability*. 1ª ed. London: Routledge (Taylor and Francis Group), 2003.
- ⁵³ Ramsey FP. *Truth and Probability* (1926). In: Kyburg HE, Smokler HE, editors. *Studies in Subjective Probability*. London: Wiley, 1964: 61-92.
- ⁵⁴ De Finetti B. *Fondamenti logici del ragionamento probabilistico* (1930). *Scritti (1926-1930)*. Padua: CEDAM, 1981: 261-263.
- ⁵⁵ De Finetti B. Probabilism (1931). *Erkenntnis* 1989; 31:169-223.
- ⁵⁶ De Finetti B. Sul significato soggettivo della probabilità. *Fundamenta mathematicae*, 17: 298-329. Reimpreso en: *On the subjective meaning of probability* (1931). *Induction and probability*. Bolonia: CLUEB, 1993: 291-321.
- ⁵⁷ Bernardo JM. Reference posterior distribution for Bayesian inference. *Journal of the Royal Statistical Society, serie B* 1979; 41(2):113-147.
- ⁵⁸ Shannon CE. A mathematical theory of communication. *Bell Systems Technical Journal* 1948; 27:379-423 y 623-656.
- ⁵⁹ Lindley DL. On a measure of information provided by an experiment. *Ann Math Statist* 1956; 27:986-1005.
- ⁶⁰ Girón FJ, Bernardo JM. El control de la incertidumbre: El cálculo de probabilidades y la teoría de la utilidad. In: Millán G, editor. *Horizontes Culturales: Las fronteras de la ciencia*. Madrid: Espasa, 2002: 205-215.
- ⁶¹ Bernardo JM. *Bioestadística: Una perspectiva bayesiana*. 1ª ed. Barcelona: Ediciones Vicens-Vives, S.A., 1981.
- ⁶² Lindley DL. The philosophy of statistics (with Discussion). *The Statistician* (Journal of the Royal Statistical Society, Series D) 2000; 49(3):293-337.
- ⁶³ Howson C, Urbach P. *Scientific reasoning: the bayesian approach*. 2nd ed. Chicago and La Salle, Illinois: Open Court, 1996.
- ⁶⁴ Keynes JM. *A Treatise on probability* (2ª ed, 1929. London, Macmillan). New York: Harper and Row, 1962.
- ⁶⁵ Carnap R. *Logical foundations of probability*. Chicago: Chicago University Press, 1950.
- ⁶⁶ Kosko B. *Pensamiento borroso (Fuzzy Thinking: the new science of fuzzy logic, 1993)*. 1ª ed. Barcelona: Crítica (Grijalbo Mondadori), 1995.

- ⁶⁷ Trillas E, Alsina C, Terricabras JM. Introducción a la lógica borrosa. 1ª ed. Barcelona: Editorial Ariel, 1995.
- ⁶⁸ Martín del Brío B, Sanz Molina A. Redes neuronales y sistemas borrosos. 2ª ed. Madrid: RA-MA Editorial, 2001.
- ⁶⁹ Lee PM. Bayesian statistics, an introduction. 2nd ed. London: Arnold Publishers, Co.(John Wiley & Sons Inc.), 1997.
- ⁷⁰ Pocock S, Spiegelhalter D. Domiciliary thrombolysis by general practitioners. *BMJ* 1992; 305:1015.
- ⁷¹ Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. *BMJ* 1999; 319(21 Agosto):508-512.
- ⁷² GREAT group. Feasibility, Safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *BMJ* 1992; 325:548-553.
- ⁷³ Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. 1ª ed. Chichester, West Sussex (England): John Wiley & sons Ltd., 2004.
- ⁷⁴ Good IJ. A M Turing's statistical work in World War II. *Biometrika* 1979; 66: 393-6
- ⁷⁵ Mahon P. Excerpt from Turing's Treatise on the Enigma. En: Copeland BJ. The essential Turing. 1ª ed. New York (USA). Oxford University Press, 2004.
- ⁷⁶ Lahoz-Beltra R. Turing: del primer ordenador a la inteligencia artificial. 1ª ed. Col: La matemática en sus personajes (Nº 24). Tres Cantos (Madrid). Editorial Nívola, 2005.
- ⁷⁷ McGee S. Simplifying Likelihood Ratios. *J Gen Intern Med* 2002; 17:647-650.
- ⁷⁸ Kardaun OJ, Salomé D, Schaafsma AGM, Willems JC, Cox DR. Reflections on fourteen cryptic issues concerning the nature of statistical inference. *International Statistical Review* 2003; 71:202-306.
- ⁷⁹ Bernardo JM. A bayesian approach to some cryptic uses on the nature of statistical inference. *International Statistical Review* 2003; 71:307-314.
- ⁸⁰ Bernardo JM. Una introducció a l'estadística bayesiana. *Butlletí de la societat catalana de matemàtiques* 2001; 17:7-64.
- ⁸¹ Bernardo JM. Bayesian statistics. In: UNESCO, editor. Volumen Probability and statistics (Viertl, R ed.) de la Encyclopedia of life support systems (EOLSS). Oxford (UK), 2003. (ww.eolss.net)
- ⁸² Berger JO, Berry DA. Statistical analysis and the Illusion of objectivity. *American Scientist* 1988; 76:159-165.
- ⁸³ Goodman SN. Toward evidence-based medical statistics. 1: The p value falacy. *Ann Intern Med* 1999; 130(12 (15 Junio)):995-1004.

-
- ⁸⁴ Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999; 130(12):1005-1013.
- ⁸⁵ Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. *Psychological Review* 1963; 70(3):193-242.
- ⁸⁶ Blackwell D, Dubins L. Merger of opinions with increasing information. *Ann Math Statist* 1962; 33:882-886.
- ⁸⁷ Bernardo JM. Noninformative priors do not exist: A discussion (with discussion). *J Statistics Planning and Inference* 1997; 65:159-189.
- ⁸⁸ Bernardo JM, Smith AFM. *Bayesian Theory*. New York (USA): John Wiley & Sons, LTD, 1994: 165-237.
- ⁸⁹ Kullback S, Leibler RA. On information and sufficiency. *The Annals of Mathematical Statistics* 1951; 22:79-86.
- ⁹⁰ Kullback S. *Information theory and statistics* (1ª Ed. 1959). 1ª ed. New York: Dover Publications (Ed original John Wiley and Sons, New York), 1968.
- ⁹¹ Savage LJ. *The foundations of statistics*. 2ª ed. New York: Dover Publications, Inc., 1972
- ⁹² Wald A. *Statistical Decision Functions*. 1ª Ed. New York. John Wiley & Sons. 1950.
- ⁹³ Von Neumann J, Morgenstern O. *Theory of Games and Economic Behaviour*. Princeton (N.J.): Princeton University Press, 1944.
- ⁹⁴ De Finetti B. Funzione caratteristica di un fenomeno aleatorio. *Memorie dell' Accademia Nazionale dei Lincei* 1930; 4:86-133.
- ⁹⁵ Bernardo JM, Smith AFM. Cap. 4: Exchangeability and related concepts. En: *Bayesian Theory*. New York (USA): John Wiley & Sons, LTD, 1994 (2ª reimpresión, Mayo 2002): 165-237.
- ⁹⁶ Bernardo JM. The concept of exchangeability and its applications. *Far East J Math Sci* 1996; 4:111-121.
- ⁹⁷ Lindley DL, Phillips D. Inference for a Bernoulli process (a Bayesian View). *The American Statistician* 1976; 30(3):112-119.
- ⁹⁸ Greenland S, Robins JM. Identifiability, Exchangeability and Epidemiological Confounding. *International Journal of Epidemiology* 1986; 15:412-418.
- ⁹⁹ Hernán MA, Robins JM, Brumback B. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association -- Applications and Case Studies* 2001; 96(554):440-448.
- ¹⁰⁰ Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11(5):550-560.
- ¹⁰¹ Hernán MA. Conocimiento experto, confusión y métodos causales. *Gaceta Sanitaria* 2001; 15(Supl 4):44-48.

- ¹⁰² Greenland S. Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias. *Epidemiology* 2003; 14:300-306.
- ¹⁰³ Lindley DL, Novick MR. The role of exchangeability in inference. *The Annals of Statistics* 1981; 9(1):45-58.
- ¹⁰⁴ Rubin DR. Bayesian inference for causal effects: The role of randomization. *Ann Statist* 1978; 6(1):34-58.
- ¹⁰⁵ Pauker SG. Clinical decision making: handling and analyzing clinical data. En: Goldman L, Bennet JC, editors. *Cecil's Textbook of Medicine*. Philadelphia (Pensilvania): WB Saunders, 2000: 76-82.
- ¹⁰⁶ Ebel MH. *Evidence-Based Diagnosis: A Handbook of Clinical Prediction Rules*. 1ª ed. New York: Springer-Verlag, 2001.
- ¹⁰⁷ Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *N Engl J Med* 1985; 313(13):793-799.
- ¹⁰⁸ El Papiro de Edwin Smith. En: López Piñero JM. *Medicina, historia, sociedad. Antología de clásicos médicos*. 3ª ed. Esplugues de Llobregat (Barcelona): Editorial Ariel, 1973.
- ¹⁰⁹ La colección hipocrática. En: López Piñero JM. *Antología de clásicos médicos*. 1ª ed. Madrid: Editorial Triacastela, 1998.
- ¹¹⁰ McGinn T, Guyatt G, Wyer P, Naylor CD, Stiell IG. Diagnosis. Clinical Prediction Rules. En: Guyatt G & Rennie D, for The Evidence-Based Medicine Working Group, editors. *User's Guides to the medical literature. A manual for evidence-based clinical practice*. Chicago, IL: AMA press, 2002: 471-483.
- ¹¹¹ Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997; 277(6):488-494.
- ¹¹² Stiell IG, McKnight RD, Greenberg GH, McDowell I, Nair RC, Wells GA et al. Implementation of the Ottawa ankle rules. *JAMA* 1994; 271(11):827-832.
- ¹¹³ Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M et al. Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA* 1993; 269(9):1127-1132.
- ¹¹⁴ Auleley GR, Ravaud P, Giraudeau B, Kerboull L, Nizard R, Massin P et al. Implementation of the Ottawa ankle rules in France. A multicenter randomized controlled trial. *JAMA* 1997; 277(24):1935-1939.
- ¹¹⁵ Mark DB, Shaw L, Harrell FE, Jr., Hlatky MA, Lee KL, Bengtson JR et al. Prognostic value of a treadmill exercise score in outpatients with suspected coronary artery disease. *N Engl J Med* 1991; 325(12):849-853.
- ¹¹⁶ Ibáñez Pradas V, Modesto i Alapont V. MBE en cirugía pediátrica. Lectura crítica de artículos. Pruebas diagnósticas (II). *Cirugía Pediátrica* 2006; 19: 130-5.

- ¹¹⁷ Lindley DL. The philosophy of statistics (with Discussion). *The Statistician (Journal of the Royal Statistical Society, Series D)* 2000; 49(3):293-337.
- ¹¹⁸ Bernardo JM. Bruno de Finetti en la estadística contemporánea. En: Ríos S, editor. *Historia de la Matemática en el Siglo XX*. Madrid: Real Academia de Ciencias exactas, físicas y naturales, 1998: 63-80.
- ¹¹⁹ Bernardo JM. Mètodes estadístics contemporanis en la investigació científica: anàlisi bayesià. *Mètode* 2000; 24:32-34
- ¹²⁰ Jaeschke R, Guyatt G, Lijmer J. Diagnostic Tests. En: Guyatt G, Rennie D (Eds). *User's guides to the medical literature*. 1ª Ed. Chicago. American Medical Association press. 2002
- ¹²¹ Gardner MJ, Altman DG (Eds). *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Journal. 1989
- ¹²² Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991; 44 (8): 763-70.
- ¹²³ Burgueño MJ, García-Bastos JL, Gonzalez-Buitrago JM. Las curvas ROC en la evaluación de las pruebas diagnósticas. *Med Clin (Barc)* 1995; 104: 661-70.
- ¹²⁴ Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.
- ¹²⁵ Lemeshow S, Hosmer D. A review of goodness of fit test for use in the development of logistic regression models. *Am J Epidemiol* 1982; 115(1): 92-106.
- ¹²⁶ Hanley, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148: 839-843.
- ¹²⁷ Pollack MM, Ruttimann UE, Geston PR. Pediatric Risk of Mortality (PRISM) score. *Crit Care Med* 1988; 16: 1110-6.
- ¹²⁸ Pollack MM. Evaluating pediatric intensive care units. En: Ayres SM, Grenvik A, Holbrook P, Shoemaker WC (editores). *Textbook of Critical Care*. 3ª edición. Philadelphia: W B Saunders Co. 1995.
- ¹²⁹ Balakrishnan G, Aitchison T, Hallworth D, Morton NS. Prospective evaluation of the Pediatric Risk of Mortality (PRISM) score. *Arch Dis Child* 1992; 67: 196-200.
- ¹³⁰ Ebel MH. *Evidence-Based Diagnosis: A handbook of clinical prediction rules*. 1ª Edición. New York: Springer-Verlag. 2001.
- ¹³¹ Málaga Diéguez I. Valoración del sistema de puntuación de riesgo de mortalidad pediátrica (PRISM) en la Unidad de Cuidados Intensivos Pediátricos del Hospital central de Asturias. Seminario de Investigación. Universidad de Oviedo, 1999.
- ¹³² Prieto Espuñes S, Medina Villanueva A, Concha Torre A, et al. Asistencia a los niños críticamente enfermos en Asturias: características y efectividad. *An Esp Pediatr* 2002; 57: 22-8.

-
- ¹³³ Málaga Diéguez I. Valoración del sistema de puntuación de riesgo de mortalidad pediátrica (PRISM) en la Unidad de Cuidados Intensivos Pediátricos del Hospital central de Asturias. Seminario de Investigación. Universidad de Oviedo, 1999.
- ¹³⁴ Prieto Espuñes S, Medina Villanueva A, Concha Torre A, et al. Asistencia a los niños críticamente enfermos en Asturias: características y efectividad. *An Esp Pediatr* 2002; 57: 22-8.
- ¹³⁵ Van Belle G. *Statistical Rules of Thumb*. 1ª Edición. New York: John Wiley & Sons, Inc. 2002.
- ¹³⁶ Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; 49: 1373-1379.
- ¹³⁷ Sociedad española de Cuidados Intensivos pediátricos. Informe técnico N° 3. Majadahonda (Madrid): Editorial Ergón. 2003.
- ¹³⁸ Gauch HG. *Scientific method in practice*. 1ª Edición. Cambridge University Press. 2003.
- ¹³⁹ Burgueño MJ, García-Bastos JL, Gonzalez-Buitrago JM. Las curvas ROC en la evaluación de las pruebas diagnósticas. *Med Clin (Barc)* 1995; 104: 661-70.
- ¹⁴⁰ Hanley JA, McNeil BJ. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.
- ¹⁴¹ DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics* 1988; 44:837-45.
- ¹⁴² Bernardo JM, Smith AFM. *Bayesian Theory*. New York (USA): John Wiley & Sons,LTD, 1994 (2ª reimpresión, Mayo 2002).
- ¹⁴³ Serrano Angulo J. *Iniciación a la estadística bayesiana*. Madrid: Ed La Muralla / Hespérides. Colección Cuadernos de Estadística. 2003.
- ¹⁴⁴ Álamo Santana F, Vázquez Polo FJ, Rodríguez Pérez JC. Herramientas para la investigación biomédica: la perspectiva bayesiana (I). *Med Clin (Barc)* 2002; 119 (7): 265-8.
- ¹⁴⁵ Álamo Santana F, Vázquez Polo FJ, Rodríguez Pérez JC. Herramientas para la investigación biomédica: la perspectiva bayesiana (II). *Med Clin (Barc)* 2002; 119 (7): 269-72.
- ¹⁴⁶ Hervada Vidal X *et al*. Epidat 3.0 programa para análisis epidemiológico de datos tabulados. *Rev. Esp. Salud Publica*, Mar./Apr. 2004, vol.78, no.2, p.277-280.
- ¹⁴⁷ Hosmer D, Lemeshow S. *Applied logistic regression*. 2ª edición. New York: Jhon Wiley & Sons, Inc. 2000.
- ¹⁴⁸ Cover TM, Thomas JA. *Elements of information theory*. 1ª edición. New York: John Wiley&Sons, Inc. 1991. pág. 18.
- ¹⁴⁹ Bernardo JM. Bayesian Statistics. En: Viertl R (ed). *Probability and Statistics*. Encyclopedia of Life Support Systems (EOLSS). Oxford (UK): UNESCO. 2003.
- ¹⁵⁰ Bernardo JM, Juarez MA. Intrinsic estimation. En: Bernardo JM, Bayarri MJ, et al. (eds). *Bayesian Statistics 7*. Oxford University Press. 2003.

- ¹⁵¹ Bernardo JM. Bayesian statistics. In: UNESCO, editor. Encyclopedia of life support systems (EOLSS). Oxford (UK): (www.eolss.net), 2003.
- ¹⁵² Jeffreys SH. Theory of Probability (3th. Ed, 1961). Oxford (England): Oxford University Press, 1998.
- ¹⁵³ Dreiseitl S, Ohno-Machado L. Logistic regresión and artificial neural network classification models: a methodology review. J Biomed Informatics 2002; 35: 352-359.
- ¹⁵⁴ Terrin N, Schmid CH, Griffith JL et al. External validity of predictive models: A comparison of logistic regresión, classification trees and neural networks. J Clin Epidemiol 2003; 56: 721-9.
- ¹⁵⁵ Berry DA. Statistics: a Bayesian perspective. Belmont (California). Duxbury Press, 1996.
- ¹⁵⁶ Albert JH. Bayesian computation using minitab. Belmont (California). Duxbury Press, 1996.
- ¹⁵⁷ Kadane JB, Seidenfeld T. Randomization in a Bayesian perspective. Journal of Statistical Planning and Inference 1990; 25: 329-45.
- ¹⁵⁸ Pérez Ransanz AR. Kuhn y el cambio científico. Ed: Fondo de Cultura Económica. 1^o Edición. 1999. México.
- ¹⁵⁹ Vandembroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med 2008; 5(3): e67. doi:10.1371/journal.pmed.0050067 (open Acces archive)
- ¹⁶⁰ D'Agostini G. Bayesian reasoning in data analysis. A critical introduction. 1^a edición. Singapur: World Scientific Publishing Co. 2003.
- ¹⁶¹ Royal RM. Statistical evidence: A likelihood paradigm. 1^a edición. London: Chapman and Hall. 1997.
- ¹⁶² Royal RM. The likelihood paradigm for Statistical Evidence. En: Taper ML, Lele SR. The nature of scientific evidence. 1^a edición. Chicago: Chicago University Press. 2004.
- ¹⁶³ Birnbaum A. Statistical methods in scientific inference. Nature 1970; 225: 1033.
- ¹⁶⁴ Mayo DG, Kruse M. Principles of inference and their consequences. En: Corfield D, Williamson J (eds) Foundations of bayesianism. 1^a edición. Dordrecht (The Netherlands): Kluwer academic publishers. 2001
- ¹⁶⁵ Mayo DG. An error-statistical philosophy of evidence. En: Taper ML, Lele SR. The nature of scientific evidence. 1^a edición. Chicago: Chicago University Press. 2004.
- ¹⁶⁶ Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman-Pearso philosophy of induction. Brit J Phil Sci 2006; 57: 323-57.
- ¹⁶⁷ Sackett DL. The tactis of performing therapeutic trials. En: Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: how to do clinical pactice research. 3^a edición. Lippincott Williams & Wilkins. 2006.
- ¹⁶⁸ Popper KR. The myth of the framework. En: Notturmo NA (ed): Defence of Science and Rationality. 1^a edición. London: Routledge. 1994.
- ¹⁶⁹ Goldman A. A causal theory of knowledge. Journal of Philosophy 1967; 64: 357-72.

-
- ¹⁷⁰ Fearn N. El filósofo en zapatillas: Las últimas respuestas a las grandes preguntas. 1ª edición. Barcelona: Ediciones Destino. Colección Imago mundi (Vol 133). 2008.
- ¹⁷¹ Stone M. The role of experimental randomization in Bayesian statistics: Finite sampling and two Bayesians. *Biometrika* 1969; 56: 681-3.
- ¹⁷² Stone M. Role of experimental randomization in Bayesian statistics: An asymptotic theory for a single bayesian. *Metrika* 1973; 20: 170-6.
- ¹⁷³ Rubin DR. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 1974; 66(5): 688-701.
- ¹⁷⁴ Kadane JB, Seidenfeld T. Randomization in a Bayesian perspective. En: Kadane JB, Schervish MJ, Seidenfeld T (eds) *Rethinking the foundations of statistics*. Cambridge University Press. 1ª edición. Cambridge studies in probability, induction and decision theory. 1999.
- ¹⁷⁵ Bossuyt PM, Reitsma JB, Bruns DE et al., STARD steering group. Towards complete and accurate reporting of diagnostic accuracy: the STARD initiative. *BMJ* 2003; 326: 41-44.
- ¹⁷⁶ Habbema JDF, Eijkemans R, Krijnen P, Knotterus JA. Analysis of data on the accuracy of diagnostic tests. En: Knotterus JA (ed) *The evidence base of Clinical Diagnosis*. London: BMJ Books. 1ª edición. 2002
- ¹⁷⁷ Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with confidence*. 2ª edición. London: BMJ Books. 2000.
- ¹⁷⁸ Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol* 1975; 12: 387-415.
- ¹⁷⁹ Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Stat Med* 2002; 21: 3093-3106.
- ¹⁸⁰ Obuchowski NA. Receiver Operating Characteristic Curves and their use in radiology. *Radiology* 2003; 229(1): 3-8.
- ¹⁸¹ Obuchowski NA, Lieber ML. Confidence bounds when the estimated ROC area is 1'01. *Acad Radiol* 2002; 9(5): 526-30.
- ¹⁸² Tilbury JB, Van Eetvelt PWJ, Garibaldi JM et al. Receiver Operating Characteristic Análisis for intelligent medical systems. A new approach for finding Confidence Intervals. *IEEE Transactions on biomedical engineering* 2000; 47 (7): 952-63.
- ¹⁸³ Obuchowski NA, Lieber ML. Confidence intervals for de receiver operating characteristic area in studies with small samples. *Acad radiol* 1998; 5(8) 561-71.
- ¹⁸⁴ Peng F, Hall J. Bayesian análisis of ROC curves using Harkov-chain Monte Carlo Methods. *Med Decis Making* 1996; 16: 404-411.
- ¹⁸⁵ Tosteson AN, Begg CB. A general regresion methodology for ROC curve estimation. *Med Decis Making* 1988; 8: 204-15.
- ¹⁸⁶ Hellmich M, Abrams KR, Jones DR y Lambert PC. A bayesian approach to a general regression model for ROC curves. *Med Decis Making* 1998; 18: 436-43.

- ¹⁸⁷ H. Jeffreys. Further significance tests. Proc. Roy. Soc. A146 (1936) 416
- ¹⁸⁸ I. Hacking, Logic of Statistical Inference. 1ª edición. Cambridge University Press. Cambridge. 1965. (pág 49).
- ¹⁸⁹ Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 1991; 44: 763-70.
- ¹⁹⁰ Savage LJ. Diagnosis and the bayesian viewpoint. En: Computer diagnosis and diagnostic methods: The proceedings of the second conference on the diagnostic process. Michigan University Press.1972.
- ¹⁹¹ Severini TH. On the relationship between bayesian and non-bayesian interval estimates. J R Statist Soc B 1991; 53(3): 611-8.
- ¹⁹² Zech G. Frequentist and Bayesian confidence intervals. EPJdirect C12; 2002: 1-81. Disponible en <http://www.edpsciences.com/articles/epjdirect/abs/2002/contents.html>
- ¹⁹³ Cousins RD. Why isn't every physicist a Bayesian? Am J Phys 1995; 63: 398.
- ¹⁹⁴ Bacallao J La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (I). Bayesianismo frente a frecuentalismo: sus respectivas implicaciones prácticas en relación con el análisis de datos. Med Clin (Barc) 1996; 107: 467-71.
- ¹⁹⁵ Bacallao J La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (II). Consideraciones críticas acerca de las pruebas de significación. Med Clin (Barc) 1996; 107: 539-43.
- ¹⁹⁶ Silva Ayçaguer L, Muñoz Villegas A. Debate sobre métodos frecuentistas vs. Bayesianos. Gac Sanit 2000; 14(6): 482-494.
- ¹⁹⁷ Davidoff F. Standing statistics right side up. Ann Intern Med. 1999;130:1019-1021
- ¹⁹⁸ Internacional Conference on Harmonisation E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. Stat Med 1999; 18: 1905-42. Disponible en: <http://www.ich.org/ich5e.html>
- ¹⁹⁹ D'Agostini G. Confidence limits: What is the problem? Is there the solution?.contribution to the Workshop on Confidence Limits, held at CERN, CERN 2000-005 (2000) 3. Disponible en: [arXiv:hep-ex/0002055]
- ²⁰⁰ Greenland S. Bayesian perspective for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 2006; 35: 765-75.
- ²⁰¹ Wittgenstein L. Tractatus logico-philosophicus. 1ª edición. Editorial Tecnos (Grupo Anaya). Madrid. 2002. pág 277 (Prosición N° 7).
- ²⁰² Agresti A, Min Y. Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2 x 2 Contingency Tables. Biometrics 2005; 61: 515-523
- ²⁰³ Mossman D , Berger J Intervals for post-test probabilities: a comparison of five methods. Medical Decision Making 2001; 498-507.

- ²⁰⁴ Bayarri, M, Berger J. The interplay between Bayesian and frequentist analysis. *Statist. Sci.* 2004; 19: 58–80.
- ²⁰⁵ Silva Ayçaguer LC. La investigación biomédica y sus laberintos. En defensa de la racionalidad para la ciencia del siglo XXI. 2009. 1ª edición. Ed Diaz de Santos.
- ²⁰⁶ Tsiatis AA. A note on goodness-of-fit test for the logistic regression models. *Biometrics* 1980; 67: 250-1.
- ²⁰⁷ Silva Ayçaguer LC. Excursión a la regression logística en ciencias de la salud. Ed Diaz de Santos. Madrid. 1ª edición. 1995
- ²⁰⁸ Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit test for the logistic regresión model. *Statistics In Medicine* 1997; 16: 965-980
- ²⁰⁹ Kuss O. Global goodness-of-fit test in logistic regresison with sparse data. *Statistics in Medicine* 2002; 21: 3789-3801.
- ²¹⁰ Farrington CP. On assessing goodness of fit of generalized linear models to spars e data. *Journal of the RoyalStatistical Society, Series B* 1996; 58(2):349–360.
- ²¹¹ Shannon CE. A matematical theory of communication. *Bell Systems Tech* 1948; 27:379-423 y 623-656.
- ²¹² Soofi ES. Capturing the intangible concept of information. *Journal of the American Statistica Association* 1994; 89(428): 1243-1254.
- ²¹³ AA. VV. Temas 36: La información. *Investigación y Ciencia (Ed española de Scientific American)*. Ed Prensa Científica SA. 2º trimestre de 2004.
- ²¹⁴ Quiroga RQ et al. Kullack-Leibler and renormalizad entropies: applications to electroencefalograms of epilepsy patients. *Phys Rev E Stat Phys Plasmas Fluids Interdiscip Topics* 2000; 62 (6 pt B). 8380-6.
- ²¹⁵ Sabesan S et al. Predictability of epileptic seizures: a comparative study using Lyapunov exponent and entropy based measures. *Biomed Sci Instrum* 2003; 39: 129-35.
- ²¹⁶ Benish WA. The use of information graphs to evaluate and compare diagnostic test. *Methods Inf Med* 2002; 41(2): 114-8.
- ²¹⁷ Benish WA. Mutual information as an index of diagnostic test performance. *Methods Inf Med* 2003; 42(3): 260-4.
- ²¹⁸ Okagaki T et al. Serie: Information, discrimination and divergence in cytology. I) Theoretical aspects and definitions. *Anal Quant Cytol Histol* 1990; 12(5): 342-7. II) Test discrimination as a measure of performance. *Acta Cytol* 1991; 35(1): 25-9. IV) Quality control in diagnostic cytology. *Acta Cytol* 1991; 35(1): 35-9. V) General symmetry of total discrimination and total divergency. *Anal Quant Cytol Histol* 1992; 14(3): 233-7. VI) Biases and errors of measurement in small samples. *Anal Quant Cytol Histol* 1992; 14(3): 238-44.
- ²¹⁹ Dragalin V, Fedorov V, Patterson S y Jones B. Kullback-Leibler divergence for evaluating bioequivalence. *Stat Med* 2003; 22: 913-30.
- ²²⁰ Lee WC. Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance. *Int J Epidemiol* 1999; 28: 521-5.

-
- ²²¹ Salvador X. Modelos causales con datos cualitativos. Una aplicación de la teoría matemática de la información. 1ª edición. Barcelona: Ed Ronsel (serie Metódica). 1997.
- ²²² Kullback S, Leibler RA. On information and sufficiency. *Ann Math Statist* 1951; 22: 79-86.
- ²²³ Kullback S. Information theory and statistics. New York: Dover. 1968 (publicado originalmente por Jhon Wilwy & Sons en 1959).
- ²²⁴ Bernardo JM, Juarez MA. Intrinsic estimation. En: Bernardo JM, Bayarri MJ, et al. (eds). *Bayesian Statistics 7*. Oxford University Press. 2003.
- ²²⁵ Metropolis N, Ulam S. The Monte Carlo method. *Journal of the American Statistical Association* 1949; 44: 335-41.
- ²²⁶ Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equations of state calculations by fast computing machine. *Journal of Chemical Physics* 1953; 21: 1087-92.
- ²²⁷ Hasings W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; 57; 97-109.
- ²²⁸ Geman S, Geman D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984; 6: 721-41.
- ²²⁹ Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5: Bayesian inference using Gibbs sampling Manual. 1996. MRC Biostatistics Unit. Institute of Public Health. Cambridge (UK).
- ²³⁰ La versión WinBUGS 1.4.3 apareció el 6 de Agosto de 2007, y se liberó para uso universal en Enero de 2009. Está disponible en: <http://www.mrc-bsu.cam.ac.uk/bugs>.
- ²³¹ A través da carta de entendemento existente entre a Consellería de Sanidade y la OPS-OMS. Está accesible libremente en Internet en la dirección <http://dxsp.sergas.es>
- ²³² Bayarri MJ, Cobo E. Una oportunidad para Bayes. *Med Clin (Barc)* 2002; 119 (7): 252-3.
- ²³³ Lucas 20, 25. Mateo 22, 21 y Marcos 12, 17. La Biblia Valenciana Interconfesional. Ed Saó. Castelló. 1996
- ²³⁴ Suchman AL, Dolan JG. Odds and likelihood ratios. In: Black ER, Bordley DR, Tape TG, Panzer RJ, editors. *Diagnosis strategies for common medical problems*. Philadelphia, PA: American College of Physicians - American Society of Internal Medicine, 1999: 31-36.
- ²³⁵ Gutierrez Cabria S. *Filosofía de la Probabilidad*. 1ª Ed. València. Tirant Lo Blanch. 1992.
- ²³⁶ Skyrms B. *Choice and Chance*. Belmont: Wadsworth Publishing Co., 1977.
- ²³⁷ De Finetti B. Foresight: its logical laws, its subjective sources (La prévision: ses lois logiques, ses sources subjectives. *Ann Inst J Poincaré* 1937/1964; 7:1-68). In: Kyburg HE, Smokler HE, editors. *Studies in subjective probability*. New York: Dover, 1980: 93-158.
- ²³⁸ Heath DL, Sudderth WD. De Finetti's theorem for exchangeable random variables. *Amer Statist* 1976; 30:333-345.
- ²³⁹ Chow YS, Teicher H. *Probability theory* (1ª Ed 1978). 2ª ed. Berlin: Springer Verlag, 1988.

- ²⁴⁰ Nyquist H. Certain factors affecting telegraph speed. *Bell System Technical Journal* 1924; April 1924: 324
- ²⁴¹ Nyquist H. Certain Topics in Telegraph Transmission Theory. *A I E E Trans* 1928; 47(April 1928):617
- ²⁴² Hartley RVL. Transmission of information. *Bell System Technical Journal* 1928; July 1928:535
- ²⁴³ Wiener N. *Cybernetics*. 1ª ed. Cambridge, Ma and New York: MIT Press and Wiley, 1948
- ²⁴⁴ Segal J. El geómetra de la información. *Investigación y Ciencia* (Edición española de *Scientific American*). Serie Temas N° 36. Barcelona. Ed Prensa Científica SA. 2º trimestre 2004.
- ²⁴⁵ Masoliver J, Wagensberg J. *Introducció a la teoria de la probabilitat i de la informació*. 1ª Edició. Barcelona. Edicions Proa. Sèrie Biblioteca Universitària N° 31. 1996.
- ²⁴⁶ Applebaum D. *Probability and information: an integrated approach*. Cambridge. 1ª Edición. Cambridge University Press. 1996
- ²⁴⁷ Bernardo JM, Juárez MA. Intrinsic estimation. *En: Bernardo JM, Bayarri MJ, et al. (eds). Bayesian Statistics 7*. Oxford University Press. 2003.
- ²⁴⁸ Hebb DO. *The organization of behaviour*. New York: H K Wiley, 1949. pp 60-78.
- ²⁴⁹ Rosenblatt F. *The Perceptrón: A Probabilistic Model for Information Storage and Organization in the Brain*. The MIT Press, 1989. pp 92-113.
- ²⁵⁰ Widrow B, Lehr MA. 10 years of Adaptive Neural Networks: Perceptron, Madaline and Backpropagation. *Proceedings of the IEEE* 1990; 78(9):1415-1442.
- ²⁵¹ Hilerá JR, Martínez VJ. *Redes neuronales artificiales. Fundamentos, modelos y aplicaciones*. 1ª ed. Madrid: Editorial RA-MA, 1995.
- ²⁵² Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 1982; 43:59-69.
- ²⁵³ Patterson D. *Artificial Neural Networks*. Singapore: Prentice Hall, 1996.
- ²⁵⁴ Haykin S. Back propagation. In: Haykin S, editor. *Neural Networks. A comprehensive foundation*. Englewood Cliffs, New Jersey: McMillan College Publishing Company Inc., 1994: 142-153.
- ²⁵⁵ Miller RA, Geissbuhler A. Clinical Diagnostic Decision Support Systems: An overview. In: Berner E, editor. *Clinical decision support systems: theory and practice*. New York: Springer, 1999.
- ²⁵⁶ Horton NJ, Switzer SS. Statistical Methods in the *Journal*. *N Engl J Med* 2005; 353(18):1977-1979.
- ²⁵⁷ Lisboa PJG. A review of evidence of health benefit from artificial neural networks in medical intervention. *Neural Networks* 2002; 15:11-39

- ²⁵⁸ Sarle WS. Neural Networks and Statistical Models. En: Proceedings of the Nineteenth Annual SAS Users Group International Conference. 1994
- ²⁵⁹ Dybowski R, Gant V. Clinical applications of artificial neural networks. Cambridge: Cambridge University Press. 2001
- ²⁶⁰ Weinstein MC, Finenberg HV. Clinical Decision Analysis. Philadelphia: Saunders Co., 1980
- ²⁶¹ Westenskow DR, Orr JA, Simon FH, Bender HJ, Frankenberger H. Intelligent alarms reduce anesthesiologist's response time to critical faults. *Anesthesiology* 1992; 77(6):1074-1079.
- ²⁶² Bishop CM. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1996.
- ²⁶³ Cathers I. Neural network assisted cardiac auscultation. *Artif Intell Med* 1995; 7(1):53-66
- ²⁶⁴ Papaloukas C, Fotiadis DI, Likas A, Mikalis LK. An ischemia detection method based on artificial neural networks. *Artif Intell Med* 2002; 24(2):167-178
- ²⁶⁵ Wang S, Ohno-Machado L, Fraser SF, Kennedy RL. Using patient-reportable clinical history factors to predict myocardial infarction. *Computers in Biology and Medicine* 2001; 31(1):1-13.
- ²⁶⁶ Ellenius J, Groth T, Lindahl B. Neural network analysis of biochemical markers for early assessment of acute myocardial infarction. *Stud Health Technol Inform* 1997; 43 Pte A:382-385
- ²⁶⁷ Zoni-Berisso M, Molini D, Viani S, Mela GS, Delfino L. Noninvasive prediction of sudden death and sustained ventricular tachycardia after acute myocardial infarction using a neural network algorithm. *Ital Heart J* 2001; 2(8):612-620
- ²⁶⁸ Djavan B, Remzi M, Zlotta A, Seitz C, Snow P, Marberger M. Novel artificial neural network for early detection of prostate cancer. *J Clin Oncol* 2006; 20(4):921-929
- ²⁶⁹ Setiono R. Extracting rules from pruned networks for breast cancer diagnosis. *Artif Intell Med* 1996; 8(1):37-51
- ²⁷⁰ Naguib R, Adams AE, Horne CH, Angus B, Smith AF, Sherbet GV et al. Prediction of nodal metastasis and prognosis in breast cancer: a neural model. *Anticancer Res* 1997; 17:2735-2741
- ²⁷¹ Birndorf NI, Pentecost JO, Coakley JR, Spackman KA. An expert system to diagnose anemia and report results directly on hematology forms. *Comput Biomed Res* 1996; 29(1):16-26
- ²⁷² McGuire WL, Clark GM. Prognostic factors and treatment decisions in axillary-node-negative breast cancer. *N Engl J Med* 1992; 326(26):1756-1761.
- ²⁷³ Bryce TJ, Dewhirst MW, Floyd CE Jr, Hars V, Brizel DM. Artificial neural network model of survival in patients treated with irradiation with and without concurrent chemotherapy for advanced carcinoma of the head and neck. *J Radiat Oncol Biol Phys* 1998; 41(2):339-345
- ²⁷⁴ Niederberg CS, Golden RM. Artificial neural networks in urology: applications, feature extraction and user implementations. In: Dybowski R, Gant V, editors. Clinical applications of artificial neural networks. Cambridge: Cambridge University Press, 2001

- ²⁷⁵ Serrano-Durba A, Serrano AJ, Magdalena JR, Martin JD, Soria E, Dominguez C et al. The use of neural networks for predicting the result of endoscopic treatment for vesico-ureteric reflux. *BJU Int* 2004; 94(1):120-122.
- ²⁷⁶ de Figueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I et al. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. *Proc Natl Acad Sci USA* 1995; 92(12):5530-5534.
- ²⁷⁷ Chow HH, Tolle KM, Roe DJ, Elsberry V, Chen H. Application of neural networks to population pharmacokinetic data analysis. *J Pharm Sci* 1997; 86(7):840-847.
- ²⁷⁸ Brier M, Zurada JM, Aronoff GR. Neural network predicted peak and trough gentamicin concentrations. *Pharm Res* 1995; 12:406-412.
- ²⁷⁹ Chen H, Chen TC, Min DI, Fischer GW, Wu YM. Prediction of tacrolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients. *Ther Drug Monit* 1999; 21:50-56.
- ²⁸⁰ Freeman R, Goodacre R, Sisson PR, Magee JG, Ward AC, Lightfoot NF. Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra. *J Med Microbiol* 1994; 40:170-173.
- ²⁸¹ Chun J, Atalan E, Kim SB, Kim HJ, Hamid ME, Trujillo ME et al. Rapid identification of streptomycetes by artificial neural network analysis of pyrolysis mass spectra. *FEMS Microbiol Lett* 1993; 114:115-119.
- ²⁸² Veselis R, Reinsel R, Sommer S, Carlon G. Use of neural network analysis to classify electroencephalographic patterns against depth of midazolam sedation in intensive care unit patients. *J Clin Monit* 1991; 7:259-267.
- ²⁸³ Zernikow B, Hostmannspoetter K, Miche E, Pielemeier W, Hornschuh F, Westerman A et al. Artificial neural network for risk assessment in preterm neonates. *Arch Dis Child Fetal Neonatal Ed* 1998; 79:F129-F134.
- ²⁸⁴ Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 1982; 79(8):2554-2558.
- ²⁸⁵ Simpson P. Foundations of Neural Networks. In: Sanchez Sinencio E, editor. *Artificial Neural Networks, Paradigms, Applications, and Hardware Implementations*. Philadelphia: Saunders, 1992: 3-27.
- ²⁸⁶ Pauker SG, Kassirer JP. Therapeutic Decision Making. *N Engl J Med* 1975; 229: 229-34
- ²⁸⁷ Kassirer JP, Pauker SG. Should the diagnostic testing be regulated?. *N Engl J Med* 1978; 293: 947-9
- ²⁸⁸ Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980; 302: 1109-17.
- ²⁸⁹ Rodríguez Artalejo F, Banegas Banegas JR, González Enríquez J, Martín Moreno JM, Villar Álvarez F. Análisis de decisiones clínicas. *Med Clin (Barc)* 1990; **94**: 348-354.
- ²⁹⁰ Latour Pérez J. *El diagnóstico*. Quaderns de Salut Pública i Administració de Serveis de Salut. 21. València. Escola Valenciana de Estudis per a la Salut. 2003.

²⁹¹ Jaffe DM, Tanz RR, Davis AT, Henretig F, Fleisher G. Antibiotic administration to treat possible occult bacteraemia in febrile children. N Engl J Med 1987;317: 1175-80.

²⁹² Fleisher GR, Rosenberg N, Vinci R, Steinberg J, et al. Intramuscular versus oral antibiotic therapy for the prevention of meningitis and other bacterial sequelae in young febrile children at risk of occult bacteraemia. J Pediatr 1994; 124: 504-12.



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante



Universitat d'Alacant
Universidad de Alicante



Reunido el Tribunal que suscribe en el día de la fecha acordó otorgar, por

a la Tesis Doctoral de Don/Dña. Vicent Modesto i Alapont la calificación de

Alicante 30 de Septiembre de 2011

El Secretario,

El Presidente,

Universitat d'Alacant
Universidad de Alicante
**UNIVERSIDAD DE ALICANTE
CEDIP**

La presente Tesis de D. _____ ha sido
registrada con el nº _____ del registro de entrada correspondiente.

Alicante ___ de _____ de _____

El Encargado del Registro,



Universitat d'Alacant
Universidad de Alicante