4 El análisis de curvas ROC

4.1 Introducción

El acrónimo ROC significa *Receiver Operating Characteristic*, aunque J. A. Swets (1986) habla de sus orígenes en Psicofísica y Teoría de la Señal, cuando denotaba "*Relative Operating Characteristic*". Lo más común hoy en día es el primer significado, sobre todo si buscamos en bases de datos bibliográficas, sea del área de conocimiento que sea. El concepto de ROC no está completo si no se añade algún otro nombre: curva ROC o análisis ROC.

La base de este análisis es en cualquier caso la curva ROC. Una curva ROC es una representación gráfica de la tasa de éxito (probabilidad de detectar correctamente una señal cuando dicha señal está efectivamente presente) frente a la tasa de falsa alarma (probabilidad de detectar una señal cuando efectivamente NO está presente) para tareas de detección con sólo dos resultados posibles (sí / no, presente / ausente), según se varía el umbral o criterio para detectar la señal a lo largo de la escala de valores a partir de los cuales se hace la detección.

El nombre *Receiver Operating Characteristic* está claramente marcado por sus orígenes en Teoría de la Señal y Psicofísica. Sus orígenes son los estudios de imágenes de radar después de 2ª Guerra Mundial. Los operadores de radar tenían que decidir si un punto que aparecía en la pantalla indicaba un elemento enemigo, amigo, o simplemente ruido. La TDS o Teoría de la Detección de Señales mide la capacidad de los operadores de radar para hacer estas distinciones (en realidad, son decisiones). Su capacidad, la de

todo el sistema incluyendo al operador, para hacerlo según se variaban las probabilidades de aparición de señales de uno u otro tipo en el fondo de ruido se denominó, dibujada como una curva, se denominó curva "Característica Operativa del Receptor", o, también, y contribuyendo a una cierta confusión sobre su nombre, "Característica Operativa Relativa".

En Psicofísica, el evento o la señal que se detectará puede ser una luz breve o un sonido en presencia de ruido, que puede estar o presente o ausente en el transcurso de un experimento típico de TDS. El observador sólo puede decir sí o no. La probabilidad de decir sí o de creer que se ha detectado la señal se puede manipular de muchas maneras, típicamente variando la intensidad entre el entorno ruidoso o variable, o variando la matriz de pagos por acertar.

En Psicofísica existen múltiples diseños experimentales propios de la TDS. Veremos brevemente el llamado 2AFC (2 Alternativas, "*Forced Choice*", o elección forzada). En Psicofísica (McNicol, 1972), durante un experimento basado en una tarea de reconocimiento con elección forzada se presentan al observador dos intervalos de estimulación. En uno de ellos siempre se produce una señal, y en el otro no, contiene ruido. En las series de ensayos, ruido y señal+ruido se asignan aleatoriamente al primer o segundo intervalo, y después de cada ensayo el observador debe señalar el intervalo que contiene la señal+ruido. En un experimento real la escala sobre la que se varía la señal varía de forma continua y puede tomar, teóricamente, un rango infinito de posibles valores.

Como podemos suponer, la TDS requiere un entorno experimental muy controlado, y precisamente, por su propia naturaleza de investigación exhaustiva de los distintos valores de la señal para determinar la capacidad de detección del observador humano en el dominio sensorial bajo estudio, requiere un conjunto enorme de ensayos. Green y Swets (1966) especifican que "típicamente" son necesarios del orden de 500 ensayos (p. 393, Apéndice III, Técnicas Experimentales) para cada condición experimental en este tipo de tareas sí/no. Por ejemplo "una función psicométrica podría estar basada en 2500 ensayos, a lo largo de dos o tres sesiones experimentales". Este hecho fue en un comienzo un condicionante importante de su aplicación a otros ámbitos, al que se refieren Hanley y McNeil (1982). Green y Swets señalan que "las pruebas clínicas

basadas en estos métodos son normalmente antiprácticas". Precisamente por esta dificultad, además de por los estrictos supuestos sobre las distribuciones de los estímulos, difícil de cumplir fuera del laboratorio, es que surgen varios enfoques alternativos, o "no paramétricos".

El nombre de J. A. Swets es esencial para comprender el desarrollo de las curvas ROC en diferentes áreas, más allá de la Psicología, o la Psicofísica, donde se desarrolló. Ya hemos visto en el capítulo 3 el papel esencial que ha representado en la expansión de estas técnicas a varias áreas de aplicación, comenzando por su aplicación pionera en el diagnóstico por la imagen. Un hito importantísimo en esta expansión es su libro, conjuntamente con Ronald M. Pickett, "*Evaluation of Diagnostic Systems. Methods from Signal Detection Theory*", en 1982.

En este capítulo nos detendremos con más detalle su aplicación en la Medicina, que es sin duda el área en que se han producido más aplicaciones y un mayor desarrollo metodológico, sobre todo por el empuje del enfoque no paramétrico.

En Medicina se utilizan de forma cotidiana los conceptos de sensibilidad y especificidad de una prueba diagnóstica. La sensibilidad es la tasa de éxito y la especificidad es 1 menos la tasa de falsa alarma. Estos conceptos son muy usados para evaluar la sensibilidad y especificidad de pruebas diagnósticas según se varía el punto de corte o umbral para que un diagnóstico sea "positivo". Desde los años 70 se están usando las curvas ROC en diferentes ámbitos de la Medicina, por su capacidad para resumir en una única medida el rendimiento o precisión o eficacia diagnóstica. Esta medida es el área bajo la curva ROC, o en sus siglas inglesas AUC (*Area Under Curve*).

Charles Metz es uno de los pioneros en la aplicación del análisis ROC al diagnóstico por la imagen con su trabajo de 1975 (Metz, Starr, Lusted y Rossman, 1975) y el artículo de 1978 en *Seminars in Nuclear Medicine*, y continúa, desde su Departamento de la Universidad de Chicago (http://xray.bsd.uchicago.edu/krl/index.htm), su trabajo tanto de desarrollo metodológico, con la publicación de un programa de ajuste de curvas ROC binormales (ROCKIT) como de divulgación (Metz, 1986).

Los otros autores que aparecen en las referencias clásicas de la aplicación del análisis ROC al campo de diagnóstico médico son Hanley y McNeil. Podemos destacar Hanley y McNeil (1982), Hanley (1989), y un artículo básico para entender el avance metodológico realizado en el campo del diagnóstico médico, el artículo clásico de Hanley y McNeil (1983) en el que proponen un método para comparar las áreas bajo las curvas ROC derivadas de los mismos casos. Otro artículo clave en este momento es el de McNeil y Hanley (1984), en el que explicitan en su título la consideración de las curvas ROC como una herramienta estadística.

En su artículo clásico aparecido en Radiology en 1982, Hanley y McNeil comentan la rápida expansión de las curvas ROC como medida "del contenido informativo de una variedad de sistemas de diagnóstico por imagen", además de su uso para comparar estadísticamente la capacidad discriminativa de métodos estadísticos, en general, siempre que se utilizaran indicadores numéricos con propósitos predictivos.

Pero además señalan las críticas que surgen hasta ese momento, recogiendo también la falta de información detallada sobre la aplicación de esta técnica. Esta última crítica viene quizá razonada por la escasa capacidad "didáctica" de muchas de las publicaciones anteriores, principalmente a cargo de Swets y Metz. Además de esta crítica, recogen algunas dificultades importantes para su aplicación más generalizada en campos aplicados:

- Las medidas cuantitativas que se utilizaban para describir una curva ROC se calculan bajo el supuesto de que "los grados variables de normalidad / anormalidad" se pueden representar por dos distribuciones gaussianas que se solapan.
- Se necesitaban métodos numéricos iterativos de estimación, más que formas algebraicas expresas, para estadísticos importantes relacionados con las curvas ROC, como el error típico del área bajo la curva. Estos métodos (se refiere sin duda a los algoritmos de ajuste de curva ROC mediante máxima verosimilitud) requieren programas de ordenador muy especiales y son difíciles de entender. Hemos de recordar que en aquella época no había ordenadores personales, y los cálculos estadísticos estaban restringidos a los grandes centros de cálculo.

- Por último, no había hasta esa fecha (1982) ningún método para estimar el tamaño de la muestra necesaria para asegurar un grado específico de precisión para un índice particular.

Hanley y McNeil describen los parámetros principales que definen el análisis ROC:

- Las curvas ROC, supuestas distribuciones normales que se superponen, se caracterizan por dos parámetros, uno es la diferencia de medias y el otro es una proporción de varianzas.
- El área bajo la curva, denominado por ellos *A(z)* -que simboliza sus raíces en el análisis de distribuciones gaussianas-, que varía entre 0.5 (no hay discriminación, se elige al azar) y 1 (discriminación perfecta, son la misma distribución gaussiana).

En ese momento, para poder calcular el error típico había que acudir a la estimación de la curva mediante máxima verosimilitud, normalmente mediante el algoritmo publicado por Dorfman y Alf (1969). Este procedimiento también permitía el cálculo de intervalos de confianza para las curvas y por tanto para las áreas bajo la curva.

¿Cómo se utiliza la curva ROC en el contexto de la evaluación de sistemas de diagnóstico por la imagen?

A diferencia del procedimiento de dos alternativas y elección forzada (2AFC) de Psicofísica -que suele exigir una cantidad impresionante de combinaciones para poder realizar las estimaciones de la curva- en los estudios de imagen médica se utiliza un método de *rating* o valoraciones en una escala simple de cinco puntos:

1 - Definitivamente normal
2 - Probablemente normal
3 - Cuestionable
4 - Probablemente anormal
5 - Definitivamente anormal

Las imágenes médicas tanto de casos enfermos como de casos sanos se entremezclan y se presentan en orden completamente aleatorio a los evaluadores (también llamados en

este contexto "lectores" o "*readers*"). Los puntos necesarios para producir la curva ROC se obtienen considerando sucesivamente categorías más amplias de anormal (por ejemplo, en una secuencia de categorías sólo 5, a continuación 5 y 4, a continuación 5, 4 y 3, ...).

Hanley y McNeil justifican que el área de la curva ROC obtenida mediante este tipo de "experimentos de valoración" (*rating experiments*) tienen el mismo significado, al menos conceptualmente, que los experimentos 2AFC de Psicofísica. Precisan, de todos modos, que esta equivalencia no lo es en el sentido estricto "o matemático", sino conceptualmente porque las curvas en la práctica se construyen de forma diferente.

El análisis ROC para la evaluación de la capacidad predictiva de instrumentos de detección rápida

Además de la evaluación de sistemas de diagnóstico por la imagen, como el caso de Radiología descrito, hay otra aplicación sobresaliente en la Medicina, la de las pruebas de detección rápida o temprana, una traducción muy complicada del sintético término inglés "*screening*". El objetivo de estas pruebas es obtener un diagnóstico rápido a partir de indicadores de fácil consecución para comenzar a trabajar (clasificar al enfermo, solicitar pruebas diagnósticas más costosas y lentas, pero con un conocimiento previo).

Un caso que puede ser interesante mostrar aquí es la tesis doctoral de Paloma Dorado (1995), quien investigó, entre otros, las pruebas Apache (II y III), que proporcionan indicadores de gravedad en una UCI. Los índices Apache II y III son indicadores numéricos de gravedad que, a partir de una determinada baremación y mediante procedimientos relativamente sencillos proporcionan una puntuación que resume la gravedad de una persona. En Medicina se insiste en que se tratan de índices para pacientes no clasificados, esto es, que no presuponen ninguna patología concreta, sino que se construyen a partir de indicadores fisiológicos válidos para cualquier persona.

Una razón del uso en la práctica de estos indicadores es como predicción de muerte. Resulta lógico suponer que a mayor gravedad, mayor probabilidad de fallecimiento. Como tales se pueden considerar herramientas de detección temprana o diagnóstico

rápido, que permiten una rápida clasificación del paciente en función de su riesgo estimado de fallecimiento.

Si representamos los intervalos de estos instrumentos frente a la probabilidad estimada de fallecimiento (denominada "*exitus*" en este área de la Medicina), a partir de los datos obtenidos con los mismos pacientes, encontraremos que una curva exponencial ajusta bastante bien la relación entre gravedad y riesgo de muerte. En la figura a continuación se pueden ver las curvas ajustadas mediante regresión logística para los indicadores clínicos Apache II y Apache III.

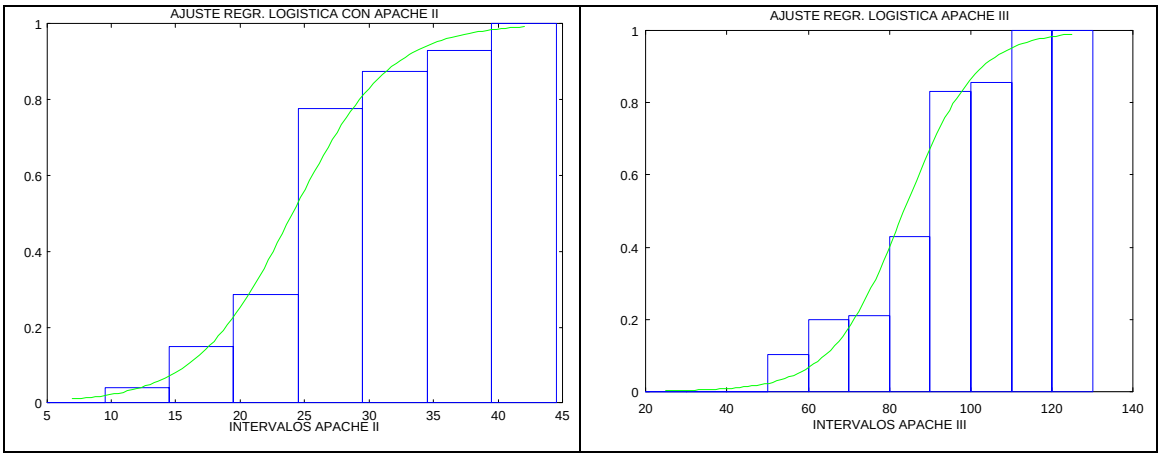

Figura 4.1. Distribuciones de dos instrumentos de detección en la UCI y su riesgo asociado

Existen varios modelos estadísticos que permiten ajustar este tipo de curvas. La curva anterior se ha obtenido mediante regresión logística. Su fórmula más básica (para un sólo predictor) es la siguiente (Hosmer y Lemeshow, 1989):

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (4.1)$$

Este tipo de estudios es muy popular en múltiples disciplinas, siempre con el objetivo de elaborar instrumentos rápidos, comprobar su eficacia, y ser capaces de compararla con otros alternativos. En el capítulo 3 se resumen muchos de estos estudios en las más variadas áreas de Medicina y en Psicología Clínica.

En este momento nos podemos plantear algunas preguntas sobre la utilidad de estos indicadores en este contexto en particular:

- ¿Qué punto o puntos de corte en la escala del indicador utilizaremos para clasificar o predecir el resultado con los diferentes enfermos? ¿Desearemos detectar a todos los que estén en riesgo de fallecimiento, incluso aunque tengamos muchos falsos positivos? ¿O por el contrario, nos puede interesar optimizar esta elección de tal manera que tengamos a la vez una cantidad razonable de detección -verdaderos positivos- pero minimizando las falsas alarmas?
- ¿Cuál de los dos indicadores tiene mejor eficacia predictiva GLOBAL? ¿Podemos tener una prueba de contraste estadístico que nos permita decidir si existen diferencias significativas entre ambas.

Utilizando los procedimientos de análisis de curvas ROC que veremos en este capítulo se pudo obtener un punto de corte maximizando la proporción entre especificidad y sensibilidad para cada prueba (situado entre 24 y 25 puntos de APACHE II, y 83 y 85 puntos de APACHE III).

Las medidas de la capacidad predictiva de herramientas de este tipo tienen siempre el gran problema de que dependen totalmente del punto de corte elegido, que depende a su vez de la escala de la prueba y de muchas otras cosas, como los costes. Las curvas ROC, una para cada instrumento, permiten la comparación visual y posteriormente estadística, y proporcionan una medida única de capacidad diagnóstica para todos los puntos de corte. Las curvas ROC empíricas presentadas por Paloma Dorado en su tesis doctoral fueron las que aparecen en la figura 4.2 a continuación.

El objetivo por tanto de este capítulo será exponer la metodología del análisis de curvas ROC para su aplicación posterior a un problema de detección en Marketing (detectar las personas que pudieran estar en riesgo de abandonar el programa por no haber intercambiado sus puntos), que se puede perfectamente asimilar a muchos problemas de detección rápida tal y como se hacen en muchos otros contextos como hemos visto en el capítulo 3.

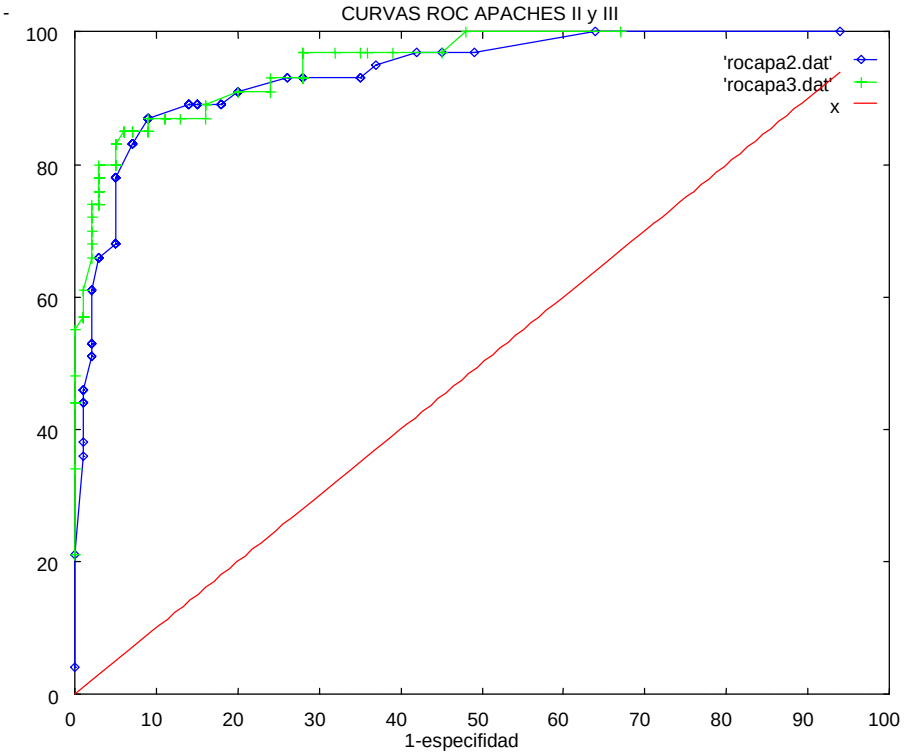

Figura 4.2. Ejemplo de curvas ROC en una prueba diagnóstica de detección rápida en la UCI.

Expondremos en primer lugar el enfoque no paramétrico, en el cual no se requiere de las distribuciones de casos positivos y negativos forma alguna, junto con las pruebas estadísticas que se pueden realizar con este enfoque. En segundo lugar, expondremos el enfoque paramétrico, que supone distribuciones gaussianas en los grupos positivo y negativo, asimismo con sus pruebas estadísticas asociadas. Una tercera parte compara ambos enfoques, y por último se presentan los desarrollos metodológicos sobre las curvas ROC, que aunque no podamos aplicar en muchos casos, muestran la expansión de este tipo de procedimientos para diferentes propósitos.

Este capítulo tiene una finalidad didáctica, que se concreta en exponer las diferentes técnicas a través de ejemplos, y aunque no evitamos las complejidades de algunos de los procedimientos, evitaremos demostraciones o prolijas justificaciones, remitiendo en cualquier caso a la literatura relevante en este tema.

4.2 Curvas ROC empíricas o no-paramétricas

4.2.1 Tipos de diseños de estudios de análisis ROC

Por razones prácticas en el desarrollo de este tema, será interesante plantear aquí los dos tipos básicos de estudios o diseños para el análisis ROC. El objeto de estos estudios será el comparar dos instrumentos o pruebas diagnósticas, con lo que tendremos que obtener dos muestras a partir de las cuales obtendremos los datos correspondientes. Al igual que con los diseños experimentales más básicos (1 sólo factor, 2 grupos) distinguiremos:

- Diseños de muestras independientes, y por tanto no correlacionados. En estos estudios los individuos positivos y negativos (p.ej., con y sin enfermedad, en riesgo o grupo control), están asignados de forma aleatoria cada uno de los instrumentos diagnósticos, de tal forma que un grupo recibe un instrumento mientras que el otro, de forma completamente independiente, recibe el otro instrumento diagnóstico
- Diseños de muestras emparejadas o correlacionadas. En estos estudios, ambos grupos reciben ambas pruebas diagnósticas.

En los procedimientos no paramétricos esta distinción tiene importantes implicaciones, por lo que será básico distinguir qué tipo de diseño es el nuestro. En múltiples aplicaciones, los diseños más comunes son los segundos: una única muestra, tanto positivos como negativos, que recibe las diferentes pruebas diagnósticas. En nuestro caso, serán también normalmente de este tipo, por el propósito de comparar la eficacia de distintos procedimientos predictivos, aunque el enfoque de "grupos independientes" se puede utilizar para comparar entre diferentes grupos (por ejemplo, el área bajo la curva ROC).

4.2.2 El punto de partida: sensibilidad y especificidad

El esquema básico de partida del análisis de curvas ROC es el mismo que el de la Teoría de Detección de Señales: la conocida tabla de doble entrada en la que se representa por un lado el resultado de una decisión, y por otro el estado real (conocido o

no). Se producen, por tanto, cuatro posibles resultados, que se representan en una tabla de doble entrada. Utilizaremos en este momento la notación y el ejemplo de Swets (1986, 1988), con la ventaja de que esta notación también es la utilizada por programas estadísticos como NCSS 2004 (Hintze, 1998, 2003). La tabla es específica para un determinado punto de corte en nuestra escala o indicador diagnóstico, por tanto, habría que hacer tantas tablas como puntos de corte en la escala, o, si se están considerando todos los valores, para todos ellos:

		Estado real		
		Positivo	Negativo	Suma
Diagnóst.	Positivo	Verdaderos Positivos (VP) a	Falsos Positivos (FP) (Falsas alarmas) b	a+b
	Negativo	Falsos Negativos (FN) (omisiones) c	Verdaderos Negativos (VN) d	c+d
	Suma	(a+c)	(b+d)	a+b+c+d

Tabla 4.1. Clasificación del resultado de una prueba diagnóstica.

Esta tabla de partida es meramente descriptiva y no tenemos por qué hacer ningún supuesto sobre la forma de la distribución, pero sí sobre los datos:

- Los datos de partida, el indicador o criterio sobre el cual se está elaborando la medida debe ser cuantitativo. Estos pueden ser estimaciones de las probabilidades, resultantes de un análisis discriminante o de una regresión logística, puntuaciones directas obtenidas en una herramienta diagnóstica o predictiva, o por puntuaciones atribuidas en una escala arbitraria que indican el «grado de convicción» que tiene un evaluador de que el sujeto pueda pertenecer a una u otra categoría.
- La variable que se intenta predecir (o de estado) es dicotómica (0 / 1). Tradicionalmente el valor 1 indica la categoría que se debe considerar positiva, a

detectar, aunque en los modernos programas informáticos se puede elegir cuál es el valor que se desea detectar.

- Se considera que los números ascendentes de la escala del evaluador o del instrumento de detección representan la creciente convicción de que el sujeto pertenece a la categoría a detectar (1). Por el contrario, los números descendentes representan la creciente convicción de que el sujeto pertenece a la otra categoría (0). En los paquetes de estadística que realizan este análisis, el usuario deberá elegir qué dirección es positiva.
- Muy importante: también se considera que se conoce la categoría real a la que pertenece el sujeto.

Esta tabla describe completamente el comportamiento de un sistema diagnóstico o de detección con dos únicos posibles resultados (positivo o negativo). Dado que, si calculamos las proporciones por columna, éstas suman 1, hay dos en vez de cuatro entradas independientes en la matriz. Por la misma razón, se deben especificar dos cantidades para obtener una representación completa del comportamiento diagnóstico. Estas dos medidas son la sensibilidad y la especificidad:

- **La sensibilidad** es la proporción de casos diagnósticados como afirmativos, a partir del criterio o regla de decisión establecido, en los que se comprueba que efectivamente sucede el estado que se pretende detectar o diagnosticar.
- **La especificidad** es la proporción de casos diagnosticados como negativos, a partir de la regla de decisión establecida, en los que se comprueba que efectivamente no sucede el estado que se pretende detectar o diagnosticar.

Al igual que los datos de partida, la sensibilidad y la especificidad se calculan para cada punto de corte o valor de la escala. Ambas medidas están relacionadas entre sí y dependen de este punto de corte, de tal manera que si la regla de decisión impuesta establece un umbral muy bajo para la decisión, tendremos alta especificidad pero baja sensibilidad, y si, por el contrario establecemos un umbral alto tendremos al contrario, alta sensibilidad pero baja especificidad.

Ambos indicadores dependen, pues, de la regla de decisión que se establezca, esto es, del umbral o punto de corte en la función o modelo estimado a partir del cual los sujetos serán diagnosticados como positivos o negativos.

A partir de esta notación en la tabla tendríamos los siguientes datos básicos. Una vez más, PARA CADA PUNTO DE CORTE:

$$\text{Sensibilidad} = \frac{a}{a+c} \tag{4.2}$$

La sensibilidad es la proporción de verdaderos positivos, esto es, clasificados como tales por la regla o punto de corte, de entre TODOS los que efectivamente lo son. Esto es, incluiremos aquí los falsos negativos, esto es, aquellos descartado por nuestro punto de corte pero que efectivamente queremos detectar:

$$\text{Sensibilidad} = \frac{Verdaderos\,Positivos}{Todos\,los\,realmente\,positivos} \tag{4.3}$$

La sensibilidad es en pocas palabras, la capacidad de un sistema para detectar un hecho, o, generalizando, la probabilidad de que el resultado de una prueba diagnóstica sea positivo cuando la enfermedad o el hecho a detectar está presente. Por eso cuando deseamos una prueba muy sensible se trata de que detecte cuantos más mejor, aunque sea a costa de detectar a muchos que efectivamente no lo sean. Cuanto más bajo el punto de corte, más sensible es la prueba. Pero esto puede no ser interesante, o puede ser sencillamente imposible de aplicar.

La otra cara de la moneda es la especificidad, definida como:

$$\text{Especificidad} = \frac{d}{b+d} \tag{4.4}$$

La especificidad es la proporción de los verdaderos negativos (descartados por nuestra regla o punto de corte) de entre todos los realmente negativos, que incluirá a los negativos que no lo son (que son efectivamente positivos):

$$\text{Especificidad} = \frac{Verdaderos\ Negativos}{Todos\ los\ realmente\ negativos} \qquad (4.5)$$

En este punto consideramos interesante poner un ejemplo más propio de Psicología, como es el instrumento de detección del maltrato propuesto por Díaz Aguado, Martínez Arias y otros (1996), que nos permitirá ilustrar todos estos conceptos en un contexto real de decisión.

4.2.3 Aplicaciones de curvas ROC descriptivas para evaluar capacidad discriminativa de una herramienta de detección temprana: el caso del maltrato

Un instrumento de detección temprana o de "*screening*" se compone de un conjunto de indicadores que proporciona una puntuación en una escala continua. El problema es el de establecer un punto de corte que nos permita predecir o detectar el objeto de nuestra detección. Este objeto puede ser, en el caso de Psicología Forense, si "reincidirá en un delito o no", o en Psicología Clínica si el individuo "tiene algún trastorno o no". Este problema es el de establecer un punto de corte o un umbral, de tal manera que nuestra decisión fuera lo más eficaz posible. La palabra "eficacia" tiene que ser definida de antemano, y dependerá mucho del contexto en que nos encontremos. Un instrumento más eficaz puede ser el que detecte al menor coste, o el que detecte lo mejor posible dejando a su vez el menor número posible de falsos negativos. Éste es por ejemplo el caso de las pruebas ELISA de detección del VIH, diseñadas para tener el mínimo porcentaje de falsos negativos, aun cuando se produzcan muchos falsos positivos, que requerirán de una segunda prueba, normalmente más costosa. Sobre este particular, véase la excelente descripción del objetivo y enfoque de estas pruebas en Swets (1992).

En este tipo de instrumentos de detección, sobre todo en casos con relevancia personal o social, el interés del investigador será detectar todos los casos en riesgo posibles, aun cuando se puedan producir muchos falsos positivos. Éste será el caso del ejemplo de la detección del maltrato infantil que trataremos a continuación.

En la investigación dirigida por Díaz Aguado y Martínez Arias (Díaz Aguado *et al*., 1996) se propone un instrumento desarrollado por los autores que permite recoger las

observaciones del profesor sobre los niños que pueden estar en situación de riesgo social, a través de un procedimiento validado y tipificado que evita las dificultades que suelen existir habitualmente en este sentido. En el análisis psicométrico se obtuvieron 4 factores como predictores del maltrato:

Factor 1: muestras visibles de maltrato físico
Factor 2: indicadores de negligencia o abandono
Factor 3: indicadores de problemas emocionales
Factor 4: observación conductas antisociales

El cuestionario piloto desarrollado constaba de 105 preguntas sobre la conducta y otros indicadores externos de cada niño, que se debía valorar en una escala de valoración de 0 a 6. Los niños sobre los que se realizó el estudio fueron previamente clasificados en dos grupos: uno de control, y otro de población en riesgo, de la que se conocía o sospechaba el maltrato. Se trata de un estudio típico de validación de un instrumento, por lo que la proporción del grupo en riesgo es muy alta (más del 30%), cuando la prevalencia del maltrato en la población es del 7%. Estos conceptos, y su repercusión sobre el análisis que estamos llevando a cabo, se estudiará más adelante.

Los datos resultantes del estudio fueron objeto de un conjunto de análisis psicométricos, que se describen brevemente a continuación. La explicación completa de estos factores y de los elementos que los componen aparece reflejada en pp. 66 y ss. de Díaz-Aguado y otros (1996).

Como base para la descripción de los resultados para cada uno de los factores en cada uno de los grupos, de tal modo que sea claramente visible el solapamiento de ambas distribuciones, resultan muy útiles los histogramas enfrentados, tal y como se muestran en las figuras 4.3 a 4.6 a continuación (véase el análisis completo en Concejero, 1998).

Para mejor ilustrar el solapamiento entre las dos distribuciones se han dibujado “en espejo”, con el histograma del grupo de riesgo por debajo del eje de abscisa.

Si comparamos la diferencia entre las distribuciones, el mayor o menor solapamiento, incluso sólo visualmente, podríamos considerar que el predictor más interesante es el

factor 2 (negligencia o abandono). Éste será un candidato obvio para nuestros análisis posteriores.

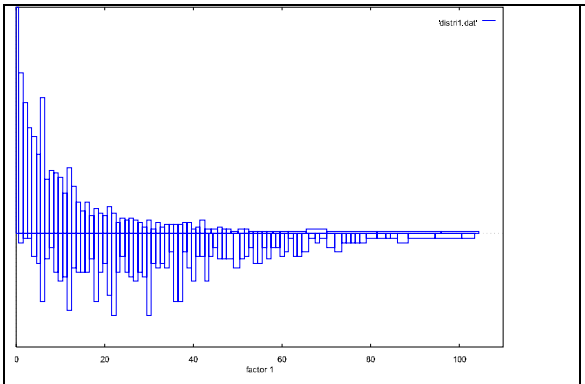

Figura 4.3. Distribución del factor 1 (Maltrato) para grupos de control (mitad superior) y riesgo (mitad inferior)

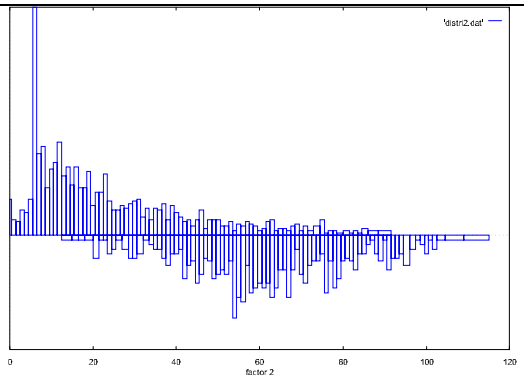

Figura 4.4. Distribución del factor 2 (Negligencia o abandono) para grupos control (mitad superior) y riesgo (mitad inferior)

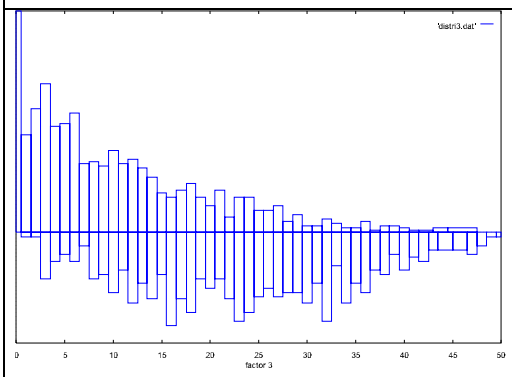

Figura 4.5. Distribución del factor 3 (Problemas emocionales) para grupos control y riesgo

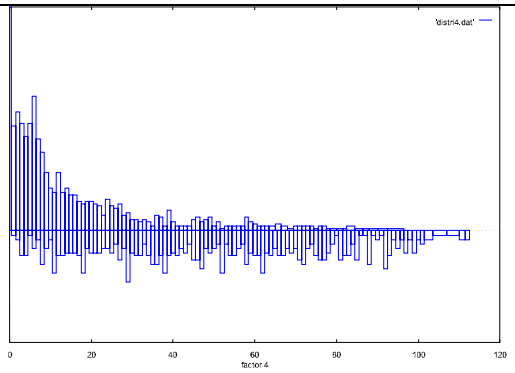

Figura 4.6. Distribución del factor 4 (Conductas antisociales) para grupos control (mitad superior) y riesgo (mitad inferior)

Además se calculó la puntuación suma del total de factores, cuyo histograma se muestra en la figura 4.7.

Podemos observar que las formas de las distribuciones, en general, no son normales, especialmente en el grupo control. Precisamente, en este tipo de pruebas, si están bien diseñadas, la variabilidad se concentra en el grupo de riesgo, permitiendo descartar a los sujetos con puntuaciones más bajas, que con una probabilidad muy alta no estarán en riesgo. Anteriormente hablamos del enfoque clásico de análisis de curva ROC, que establecía el supuesto de distribuciones normales para los dos grupos. Observamos aquí cómo este supuesto difícilmente lo podremos cumplir, puesto que las formas de las

distribuciones son muy diferentes, y no llegaríamos a una transformación que nos permitiera convertir las dos distribuciones en normales.

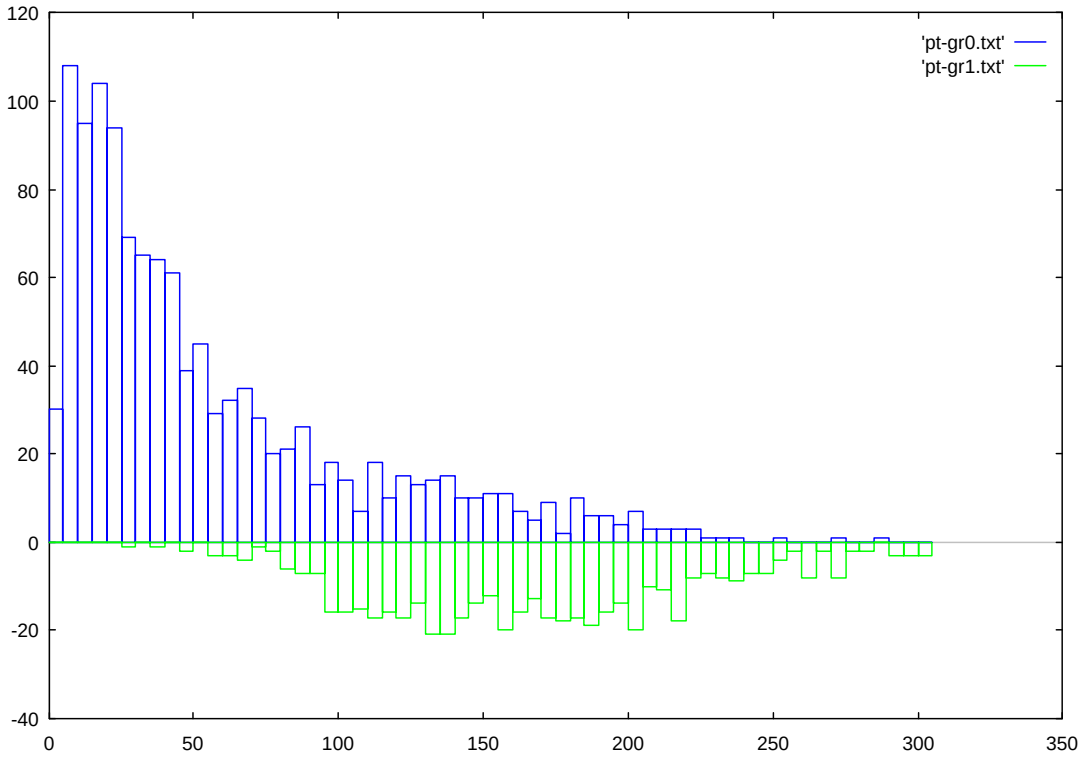

Figura 4.7. Distribuciones de puntuaciones para la puntuación de todos los factores menos el 2 para grupo control (por encima del eje de abscisa) y para grupo riesgo (por debajo.

Tomemos por ejemplo el caso del factor 2. En esta subescala la distribución para el grupo 1 (en riesgo) es normal (Shapiro-Wilks=0,994622, p= 0.0625; Kolmogorov-Smirnov= 0.0332, no significativo), mientras que la distribución del grupo control no es normal bajo ningún criterio, está claramente sesgada hacia la izquierda (valores más bajos). Por tanto, una transformación común para ambas distribuciones haría que una de ellas dejase de ser normal.

Podríamos pensar que si eliminásemos casos extremos del grupo de control, que evidentemente existen (véase la figura 4.8), podríamos acercar esta distribución al menos a una distribución simétrica. Pero la razón de ser de un instrumento de detección es que se pueda aplicar a todos los casos, sin que existan, al menos conceptualmente, casos extremos, que en todo caso serán casos "en riesgo". Por tanto, en estos casos no

podríamos llevar a cabo el análisis según el enfoque de análisis ROC tradicional, y por eso usaremos el modelo no paramétrico.

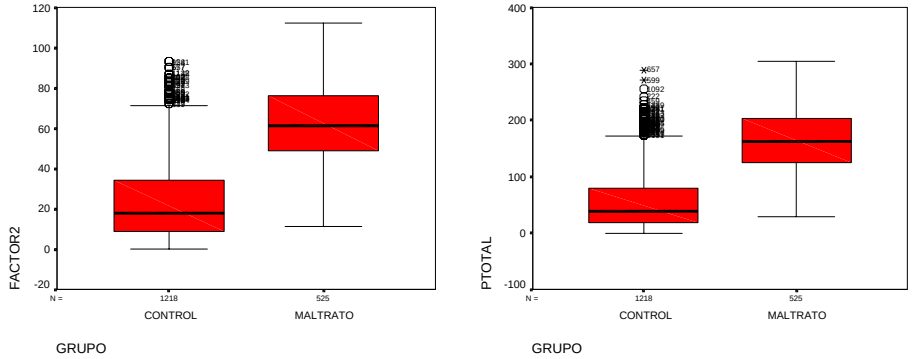

Figura 4.8. Gráficos de caja y bigotes para la subescala de abandono (factor 2) y para la puntuación total en los grupos de control y riesgo

Una vez que disponemos de los datos nos pueden surgir las siguientes preguntas:

- ¿A partir de qué puntuación en nuestra escala de predicción diremos que un niño puede estar en situación de riesgo?
- ¿Cuál es la capacidad predictiva de cada factor que compone la prueba? ¿Cuál es que mejor predice, y si tiene diferencias significativas con los restantes? ¿Necesitaríamos los cuatro factores para tener una capacidad predictiva aceptable, o podemos utilizar uno sólo el factor 2, o necesitamos además de él los restantes factores?

Ambas preguntas están íntimamente relacionadas. Esto es, la eficacia para detectar elementos de un grupo u otro dependerá totalmente del punto o valor de la escala correspondiente a partir del cual tomaremos una decisión u otra. Esto es, un punto de corte o umbral.

Tomemos por ejemplo dos intervalos en los que tenemos una alta sensibilidad: en el caso del factor de abandono será el intervalo entre 45 y 50 puntos. Si realizamos la tabulación cruzada de las puntuaciones reales obtenidas en la prueba y el resultado de utilizar cada uno de los valores como punto de corte, de tal manera que por encima de ella diremos que los sujetos están en el grupo de riesgo, tendremos, para ese preciso punto de corte, un valor de sensibilidad y especificidad como hemos definido

anteriormente. La tabla 4.2 presenta estos resultados para el factor de abandono y la tabla 4.3 para la puntuación suma del resto de factores para los intervalos de puntuación tomados.

Tabla 4.2. Resultados de sensibilidad y especificidad para el intervalo entre 40 y 50 puntos del factor de abandono, o factor 2 del instrumento de detección de maltrato

Punt.	VP a	FP b	FN c	VN d	Sensibilidad	Especificidad
41.00	457	221	68	997	0.87048	0.81856
42.00	454	211	71	1007	0.86476	0.82677
43.00	444	204	81	1014	0.84571	0.83251
44.00	437	196	88	1022	0.83238	0.83908
45.00	431	191	94	1027	0.82095	0.84319
46.00	420	183	105	1035	0.80000	0.84975
47.00	411	170	114	1048	0.78286	0.86043
48.00	406	166	119	1052	0.77333	0.86371
49.00	400	158	125	1060	0.76190	0.87028
50.00	390	150	135	1068	0.74286	0.87685

Tabla 4.3. Resultados de sensibilidad y especificidad para el intervalo entre 105 y 115 puntos de la puntuación suma de factores 1,3 y 4.

Punt.	VP a	FP b	FN c	VN d	Sensibilidad	Especificidad
105.00	463	210	62	1008	0.88190	0.82759
106.00	456	208	69	1010	0.86857	0.82923
107.00	453	207	72	1011	0.86286	0.83005
108.00	450	204	75	1014	0.85714	0.83251
109.00	448	204	77	1014	0.85333	0.83251
110.00	443	203	82	1015	0.84381	0.83333
111.00	441	201	84	1017	0.84000	0.83498
112.00	437	198	88	1020	0.83238	0.83744
113.00	433	196	92	1022	0.82476	0.83908
114.00	431	191	94	1027	0.82095	0.84319
115.00	428	186	97	1032	0.81524	0.84729

Se trata por tanto de establecer puntos de corte con un objetivo pre-determinado. Los individuos con puntuaciones por encima del punto de corte se clasifican como negativos y los que caen por debajo como positivos. Para cada punto, por tanto, tendremos un número o porcentaje de clasificados correctamente (identificados positivos que realmente lo son), o verdaderos positivos, y falsos positivos (identificados positivos que realmente no lo son), además de los verdaderos negativos y falsos negativos. Estos valores configuran una tabla 2x2 de observados frente a predichos.

Siguiendo con nuestro ejemplo, tendríamos una tabla de doble entrada para cada punto de corte. Esto es, para cada valor de la puntuación total en la escala de detección de maltrato, tendríamos un resultado de la asignación a cada grupo en función de dicho punto de corte. Por ejemplo, para el valor 45 en el factor 2:

Tabla 4.4. Resultado de la clasificación en grupos de control y riesgo para el punto de corte 45 del factor 2. Frecuencias

		Estado real		
		Forma parte del grupo de maltrato	No forma parte de grupo maltrato (grupo control)	Suma
Diagnóstico (punto corte F2=45)	Positivo (se estima que es de grupo maltrato)	Verdaderos Positivos (VP) a= 431	Falsos Positivos (FP) (Falsas alarmas) b= 191	a+b= 622
	Negativo (se estima que es de grupo control)	Falsos Negativos (FN) (omisiones) c= 94	Verdaderos Negativos (VN) d= 1027	c+d= 1121
	Suma	a+c= 525	b+d= 1218	a+b+c+d=1743

Esta tabla en proporciones por columnas aparece en la tabla 4.5 a continuación. La casilla señalada en azul es la sensibilidad de la prueba, y la verde la especificidad, ambos, es importante destacarlo, *para ese punto de corte*. En este ejemplo podemos ver cómo conseguimos una sensibilidad alta a costa de una especificidad mediana, esto es, tendremos un número de falsas alarmas bastante alto.

En la tabla 4.6, y sobre fondo amarillo, aparece la proporción total de nuestra muestra que realmente es positivo. Si fuera nuestra población, esta proporción sería la

prevalencia o tasa base de nuestro hecho para detección. Vemos cómo en nuestro caso, esta muestra "sobrerrepresenta" los individuos en riesgo, puesto que la prevalencia para maltrato infantil en la población es del 7%.

Tabla 4.5. Resultado de la clasificación en grupos de control y riesgo para el punto de corte 45 del factor 2. Porcentajes por columna

		Estado real		
		Forma parte del grupo de maltrato	No forma parte de grupo maltrato (grupo control)	
Diagnóstico (punto corte F2=45)	Positivo (se estima que es de grupo maltrato)	**Verdaderos Positivos (VP)** **0.8210**	**Falsos Positivos (FP) (Falsas alarmas)** **0.1568**	0.610
	Negativo (se estima que es de grupo control)	Falsos Negativos (FN) (omisiones) 0.1790	Verdaderos Negativos (VN) 0.8432	0.390
	Suma	1	1	1

La tabla 4.6 muestra las proporciones por filas. En esta tabla observamos cómo el total por columnas nos da un dato importante: la proporción de casos realmente positivos que hay en nuestra muestra. En nuestro caso se trata del 30% de la muestra. ¿Es esta proporción equivalente a la probabilidad de tener un caso verdaderamente positivo en la población? En el caso del maltrato infantil, la prevalencia (así se conoce este concepto) en la población es del 7%, por tanto estamos sobrerrepresentando la población en riesgo. Este dato será importante para después estimar correctamente algunos indicadores de eficacia.

Tabla 4.6. Resultado de la clasificación en grupos de control y riesgo para el punto de corte 45 del factor 2. Porcentajes por fila.

		Estado real		
		Forma parte del grupo de maltrato	No forma parte de grupo maltrato (grupo control)	Suma
Diagnóstico	Positivo (se estima que es de grupo maltrato)	Verdaderos Positivos (VP) 0.6929	Falsos Positivos (FP) (Falsas alarmas) 0.3071	1
	Negativo (se estima que es de grupo control)	Falsos Negativos (FN) (omisiones) 0.0839	Verdaderos Negativos (VN) 0.9161	1
		0.301	0.699	1

En resumen, la sensibilidad es en pocas palabras, la capacidad de un sistema para detectar un hecho, o, generalizando, la probabilidad de que el resultado de una prueba diagnóstica sea positivo cuando la enfermedad o el hecho a detectar está presente. Por eso cuando deseamos una prueba muy sensible se trata de que detecte cuantos más mejor, aunque sea a costa de detectar a muchos que efectivamente no lo sean. Como se puede ver en la tabla cuanto más bajo el punto de corte, más sensible es la prueba. Pero esto puede no ser interesante, o puede ser sencillamente imposible de aplicar.

La especificidad es la capacidad del sistema para descartar los hechos que no nos interesan, que no son objeto de detección, o, generalizando, estimaría la probabilidad de que el resultado de una prueba diagnóstica sea negativo cuando la enfermedad no está presente. Cuanto más bajo el punto de corte, MENOR nuestra especificidad.

Todas estas medidas, aunque dependientes del punto de corte, sí que permiten establecer un entorno que permite comparar la eficacia de diferentes puntos de corte, o en el caso de diferentes pruebas, también de su eficacia, puesto que no son medidas dependientes de ninguna escala o variabilidad.

Aún quedarían dos relaciones más, elaboradas a partir de las anteriores:

$$\text{Razón de verosimilitud positiva} = \frac{\text{sensibilidad}}{1 - \text{especificidad}} \qquad (4.6)$$

La razón de verosimilitud positiva mide el valor de la prueba para incrementar la certeza sobre un diagnóstico positivo (Hintze, 1998, 2003) en un punto dado. También depende de la elección de punto de corte, como todos los índices anteriores, y en el caso del punto de corte elegido para el factor 2 (45 puntos), sería 5.23.

Los valores que toma varían entre 1 e infinito, y muestra una función creciente que para el factor 2 del instrumento de detección de maltrato sería:

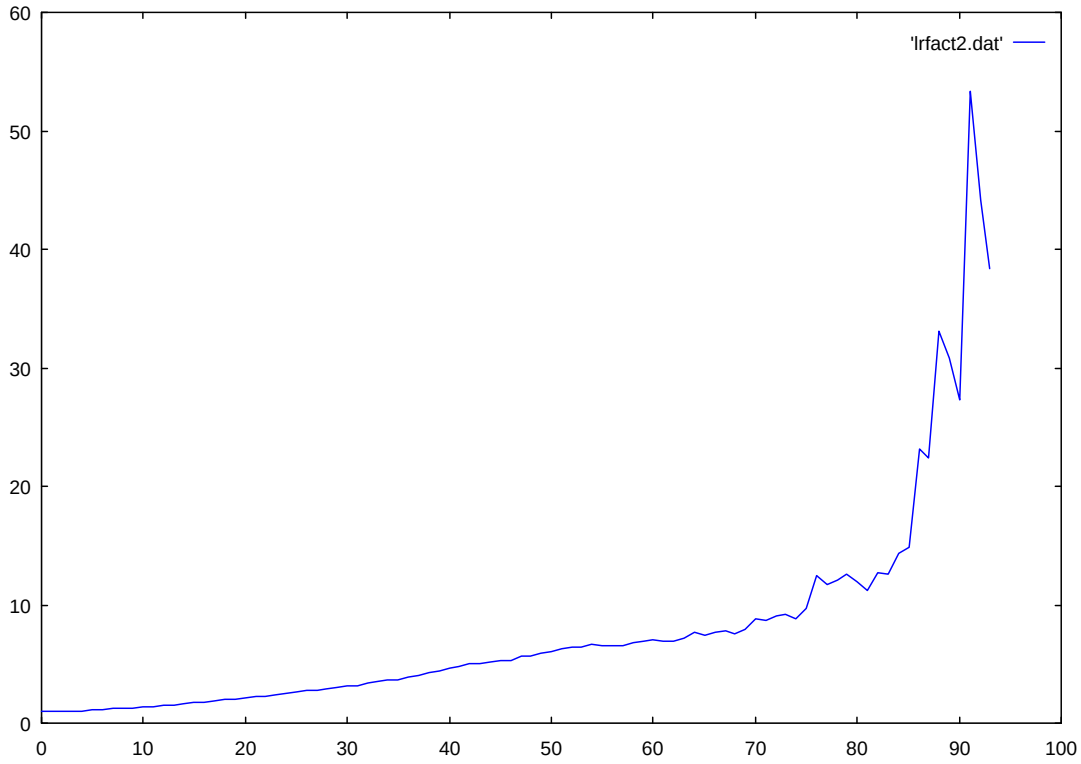

Figura 4.9. Función de la razón de verosimilitud positiva para la subescala de abandono (factor 2)

También podríamos definir su inversa, o razón de verosimilitud negativa, que mediría el valor de la prueba en este caso para incrementear la certeza sobre un diagnóstico negativo para un punto dado. Su utilidad vendrá dada por el contexto en que se utilice la prueba.

$$\text{Razón de verosimilitud negativa} = \frac{1 - \text{sensibilidad}}{\text{especificidad}} \qquad (4.7)$$

En el caso del punto de corte para el factor 2, 45 puntos, el valor de razón de verosimilitud negativa será 0.21.

Este indicador se mueve entre 0 y 1. Para el caso del factor 2 de maltrato tendría la siguiente función:

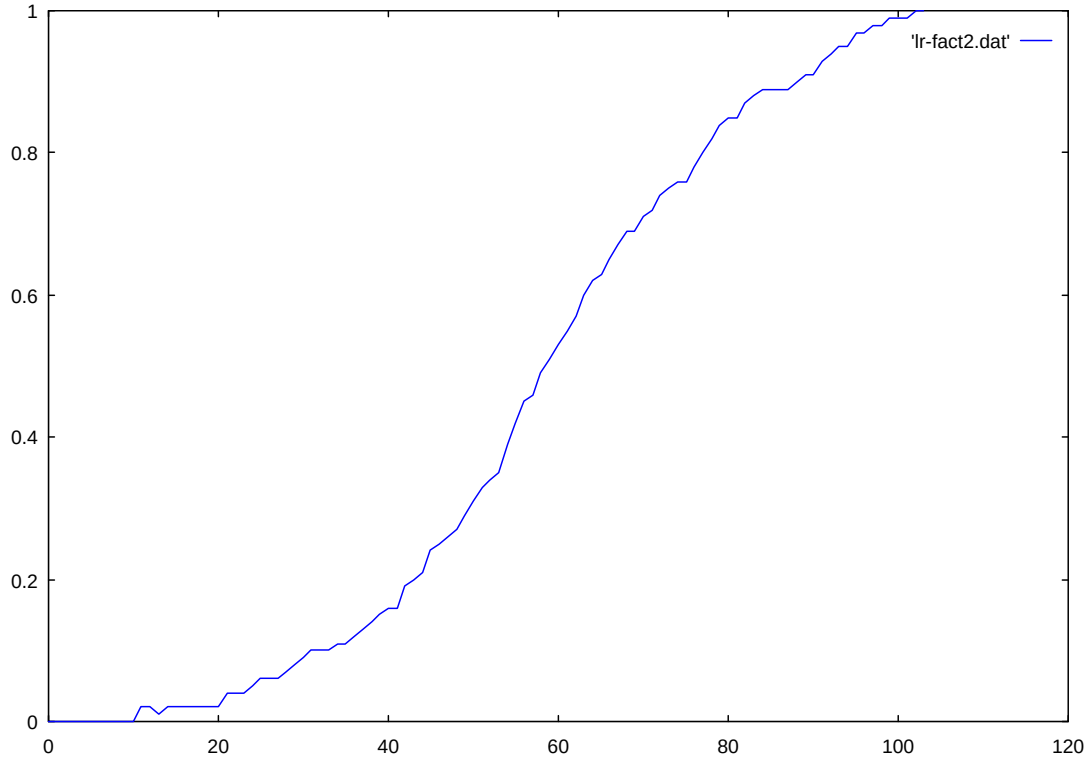

Figura 4.10. Función de la razón de verosimilitud negativa para la subescala de abandono (factor 2)

4.2.4 Cómo determinar un punto de corte óptimo

En general, nos será muy difícil acordar qué entendemos por "punto de corte óptimo", que en cualquier caso requerirá establecer previamente los objetivos tanto en eficacia de nuestro diagnóstico como en los costes que ello representa. Comúnmente podríamos considerar el mejor punto aquél que consigue la mejor relación sensibilidad y especificidad, y a la vez que minimice el coste promedio. Pero esto depende enormemente del contexto, y en muchos casos será simplemente imposible conocer este

coste. Esta definición no es nada trivial, y en realidad hay todo un campo de estudio para intentar determinar estos "costes" (Hintze, 2003), que pueden ser económicos o no.

Una de las primeras preguntas es nuestro objetivo con la decisión, y establecer nuestro enfoque que puede ser muy estricto o conservador, muy liberal, o intermedio (Swets, 1992). El primero es el típico de las pruebas de contraste de hipótesis, o del enfoque de control de calidad industrial. Se trata de establecer un punto de corte que nos proporcione una prob. de falsos positivos de entre 0.05 y 0.01, típicamente. El criterio más liberal se establece para que la probabilidad de falsos positivos sea alta, y deriva de un enfoque de ingeniería denominado "resistencia a fallos". Swets pone el ejemplo de algunos sistemas militares, cuyo umbral es relativamente bajo, de tal modo que la probabilidad de una omisión sea mínima. Éste es también el enfoque de algunas técnicas de detección, en particular la del ELISA que se ha mencionado en alguna ocasión. El criterio intermedio es el de encontrar el mejor equilibrio entre especificidad y sensibilidad, y es el más comúnmente utilizado en las técnicas de detección clínica que hemos visto en el capítulo 3. Swets se refiere a este enfoque como "simétrico", porque en el fondo, considera los costes y beneficios de cada uno de los resultados de igual modo. Se trata por tanto del enfoque más simple.

Este enfoque es el que proporciona el programa BMDP-LR (Dixon, 1992). En la tabla de doble entrada, tendríamos la siguiente clasificación de beneficios y costes (adaptando la notación original de BMDP a nuestra notación), y usaríamos la siguiente fórmula basada en el coste o "pérdida" (*loss*):

$$LOSS = k_1 a + k_2 b + k_3 c + k_4 d \qquad (4.8)$$

- El resultado de a sería normalmente 0 o positivo en cuanto a costes, puesto que representaría el beneficio por clasificar un caso como positivo cuando realmente es positivo. BMDP toma por defecto $k_1=C_{VP}=0$.
- El de b sería sería negativo: sería el gasto por producir una falsa alarma. BMDP asigna por defecto $k_2= C_{FP}= (-1)$.
- El resulado de c sería negativo: el gasto producto de omitir o no clasificar correctamente un positivo que sí lo es. BMDP asigna por defecto $k_3= C_{FN}= (-1)$.

- El resultado de d sería positivo: el beneficio por clasificar correctamente un negativo cuando lo es. BMDP asigna por defecto $k_4 = C_{VN} = 0$.

Por tanto, y si no le especificamos otros coeficientes k_1,..., k_4, la función por defecto es

$$LOSS = -(b + c) \qquad (4.9)$$

Se tratará entonces de elegir el punto con el menor coste o pérdida. Podremos fácilmente calcular esta función para nuestros indicadores, y representarlos luego para intervalos específicos en consideración. Observamos entonces cómo un punto que según esta sencilla regla minimiza el coste total para el factor 2 es 42, aunque hay diferencias muy pequeñas entre los puntos cercanos, y de hecho puede haber varios valores con el mínimo coste. En el caso de la puntuación total el valor que proporciona un máximo (o coste mínimo) es 102. En las figuras 4.11 y 4.12 se representan estas funciones de coste para dos intervalos de los indicadores de maltrato, para el factor 2 y para la suma de factores respectivamente, suponiendo, como hemos visto un coste igual a las diagnosis incorrectas en cada grupo.

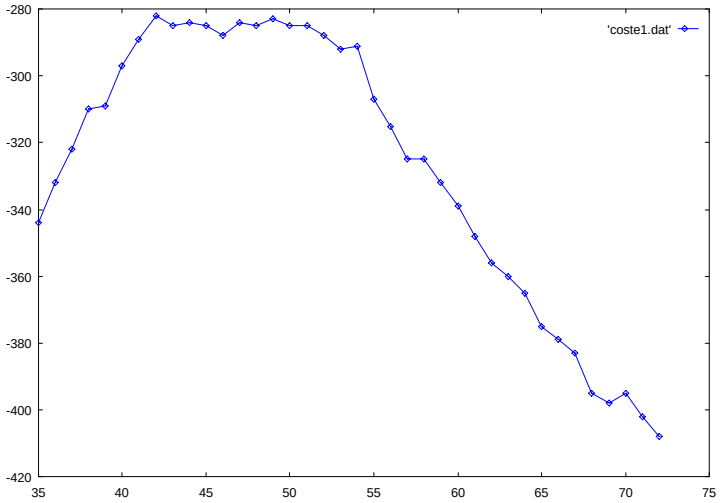

Figura 4.11. Función (simple) de coste de la clasificación en función del punto de corte del factor 2 del instrumento de detección de maltrato

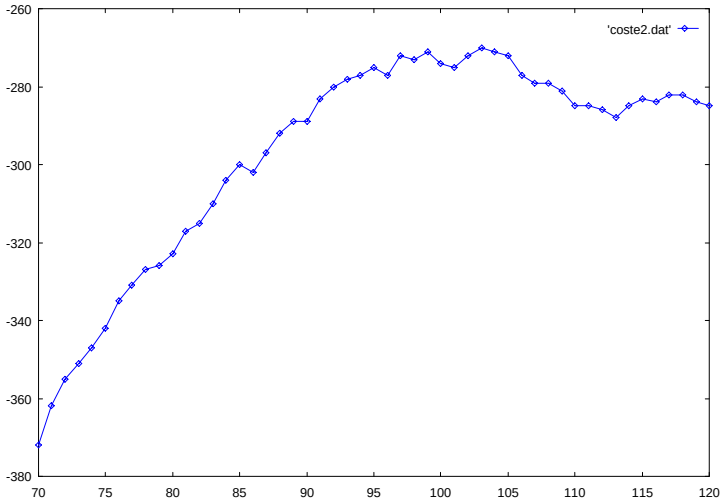

Figura 4.12. Función (simple) de coste de la clasificación en función del punto de corte de la puntuación total del instrumento de detección de maltrato

Metz (1978), recogido en Zhou *et al*. (2002) y Hintze (2003), añade el coste fijo de realizar la prueba, además de los costes y beneficios de los cuatro posibles resultados según están definidos en la tabla de doble entrada. Una vez se han estimado estos costes, el coste medio de la prueba en su conjunto para un punto de corte sería:

$$C = C_0 + C_{VP}P(VP) + C_{VN}P(VN) + C_{FP}P(FP) + C_{FN}P(FN) \qquad (4.10)$$

En esta fórmula, C_0 es el coste fijo de realizar la prueba, C_{VP} el coste asociado con un verdadero positivo, P(VP) es la proporción de verdaderos positivos en la población, y así sucesivamente.

Metz (1978) mostró que el punto donde el coste promedio es mínimo es el punto donde:

$$\text{sensibilidad} - m\,(1 - \text{especificidad}) \text{ es máximo} \qquad (4.11)$$

Donde a su vez

$$m = \frac{P(condición = falso)}{P(condición = verdadero)}\left(\frac{C_{FP} - C_{VN}}{C_{FN} - C_{VP}}\right) \qquad (4.12)$$

P(condición=verdadero) es la prevalencia del objeto de detección en la población. Dependiendo del método que se haya usado para obtener la muestra, puede o puede no ser posible estimarlo a partir de la muestra.

El programa informático NCSS 2004 (Hintze, 2003) realiza este cálculo de la siguiente forma: en primer lugar nos solicita introducir la prevalencia del caso positivo en la población (si es conocida; también admite introducir rangos) y la relación de costes según la proporción en la fórmula 4.12. Asimismo también se pueden establecer rangos para poder hacer comparaciones con varias proporciones de costes. Con el ejemplo del factor 2 que estamos tratando, podemos probar a introducir la prevalencia de 0.07, y como valores de razón coste-beneficio 0.5, 1.0 y 1.5. El resultado de esa función está representado en la figura 4.13 que aparece a continuación. Observamos cómo esta función no llega a un mínimo en un rango razonable para nuestra prueba. Se trata de decidir el punto de corte a partir del cual los beneficios no se incrementan significativamente. Aun así, según es mayor la razón de costes, mayor será el punto de corte a partir del cual no se produzca este incremento significativo. Se trata pues de una de las decisiones más importantes para el analista para la elección de un punto de corte, para lo cual debiera disponer de estimaciones fiables de costes.

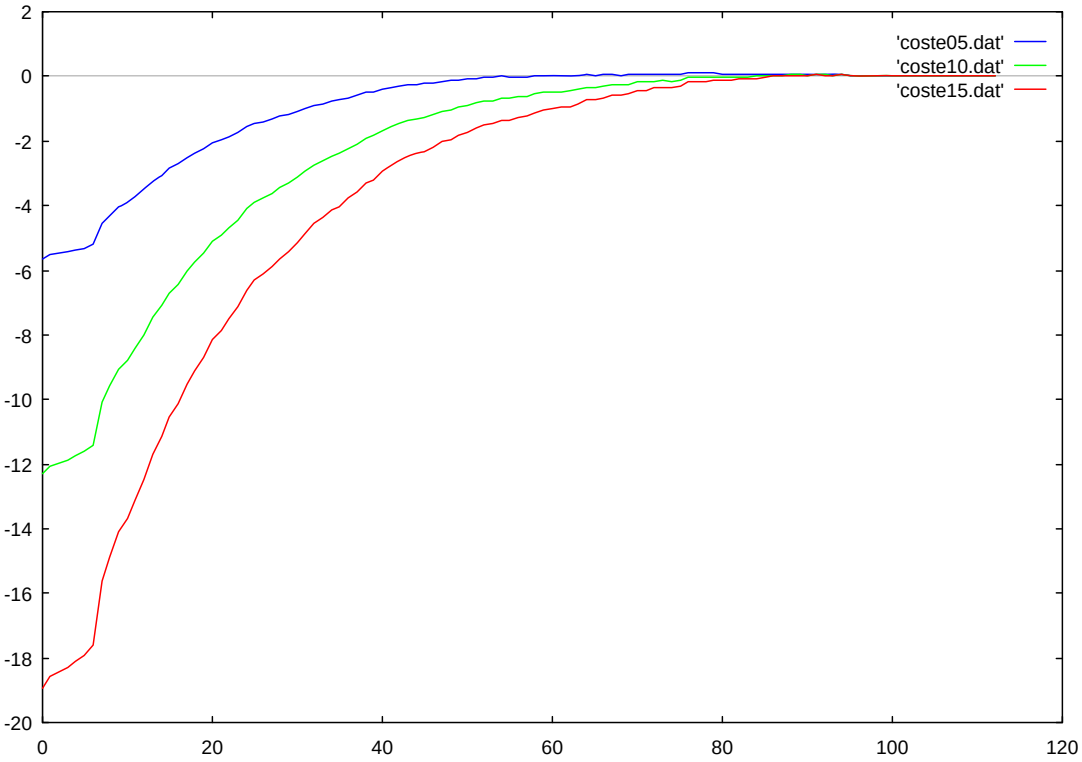

Figura 4.13. Funciones de coste según enfoque de Metz y NCSS 2004 de la clasificación en función del punto de corte del factor 2

Swets considera esta aproximación mediante razones, y no mediante la estimación de cada uno de los componentes de la ecuación 4.12, un enfoque demasiado simple, aunque pueda servir como una aproximación inicial. Efectivamente, la razón puede servir en los casos en los que no hay estimaciones de los costes, pero en rigor se debieran estimar cada uno por separado, que es la clave de este análisis.

4.2.5 La importancia de la tasa de prevalencia

La prevalencia es un concepto esencial: la prevalencia es la proporción total de individuos que son realmente positivos en la población total. Es importante señalar que este valor en ningún caso depende del punto de corte en nuestra prueba. Aun cuando tuviéramos toda la población en nuestra talba de doble entrada, la prevalencia vendría definida en término de valores marginales de la tabla de clasificación 2x2.

Un problema con la sensibilidad y la especificidad es que, por sí mismos, no evalúan la probabilidad de realizar un diagnóstico correcto. Para solucionar esto, los que utilizan este esquema en la práctica han desarrollado otros dos índices, que son el valor predictivo positivo, o PPV, y el valor predictivo negativo, o NPV (véase Hintze, 2003). Desafortunadamente, estos indicadores tienen la desventaja de que dependen directamente de la tasa de prevalencia. Por ejemplo, si el procedimiento de muestreo se construye para obtener más individuos positivos que en la población completa de interés, los valores PPV y NPV deben ser ajustados a la probabilidad real en la población.

Utilizando el teorema de Bayes, los valores ajustados de PPV y NPV se calculan basándose en los nuevos valores de prevalencia según sigue:

$$PPV = \frac{\text{sensibilidad x prevalencia}}{\left[\text{sensibilidad x prevalencia}\right] + \left[(1 - \text{especificidad}) \text{ x } (1 - \text{prevalencia})\right]} \qquad (4.13)$$

$$NPV = \frac{\text{especificidad x } (1 - \text{prevalencia})}{\left[(1 - \text{sensibilidad}) \text{ x prevalencia}\right] + \left[\text{especificidad x } (1 - \text{prevalencia})\right]} \qquad (4.14)$$

Mientras que la sensibilidad y la especificidad, y por tanto la curva ROC, y las razones de verosimilitud positiva y negativa, son todos independientes de la tasa de prevalencia de la enfermedad, los valores predictivos positivos y negativos dependen mucho de las proporciones de sujetos que tienen o no tienen realmente la enfermedad o el hecho que queremos detectar. Éstas son las probabilidades a priori de la enfermedad, o tasa base de prevalencia. Clínicamente, la prevalencia de una enfermedad es la probabilidad a priori de que el caso o el sujeto sea realmente positivo antes de que el se lleve a cabo la prueba diagnóstica. Los valores de PPV y 1-NPV son las probabilidades a posteriori de que un sujeto sea positivo después de que se haya realizado la prueba.

Si los tamaños de las muestras del grupo positivo y negativo no se corresponden con la prevalencia real de la enfermedad, hay que indicarlo en los programas estadísticos puesto que los valores predictivos positivo y negativo no tendrán sentido. Y si se conoce la prevalencia en la población, se puede entrar este porcentaje en una caja especial para que se tenga en cuenta.

Las razones de verosimilitud positiva y negativa también se deben interpretar con cuidado porque se malinterpretan fácilmente y sucede con frecuencia. Esto datos sirven para calcular los valores predictivos positivo y negativo, si se conoce la tasa de prevalencia, utilizando el teorema de Bayes.

Como ejemplo de la importancia de la tasa de prevalencia para evaluar la capacidad predictiva, tomemos el factor 2 del instrumento de detección del maltrato, con 3 tasas de prevalencia: 0.05, 0.07 y 0.10:

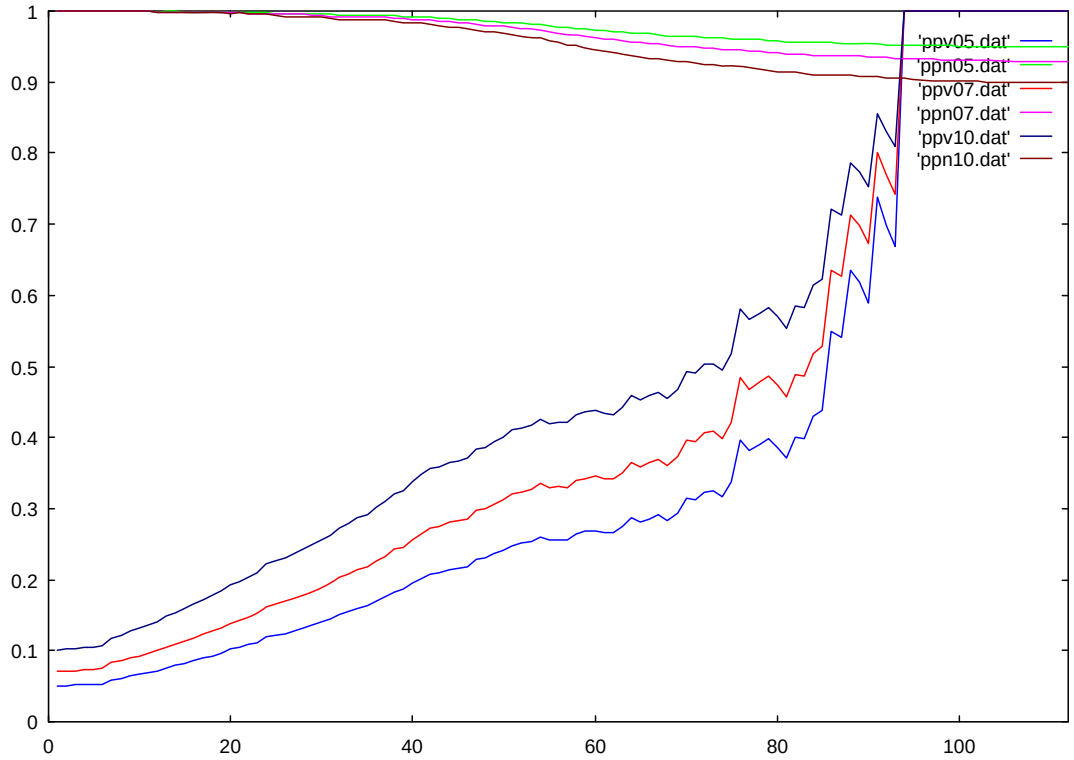

Figura 4.14. Valores predictivo positivo y negativo de la escala de abandono o factor 2, para tres tasas de prevalencia de maltrato infantil en la población: 0.05, 0.07 y 0.10. Las funciones crecientes son el valor predictivo positivo, y las decrecientes el negativo. A su vez, la creciente inferior corresponde a la prevalencia de 0.05, la siguiente a 0.07 y así sucesivamente

Observamos cómo según disminuye la tasa base, más difícil es maximizar el valor predictivo positivo, o la medida de la capacidad de la prueba para detectar los casos positivos de la población.

4.2.6 La relación entre sensibilidad y especificidad: curva ROC

Ahora podríamos dibujar el gráfico que nos relacionara cada punto de corte y los datos básicos de la prueba que estamos manejando, sensibilidad y especificidad. Para nuestras escalas de maltrato (factor 2 y puntuación total) serían:

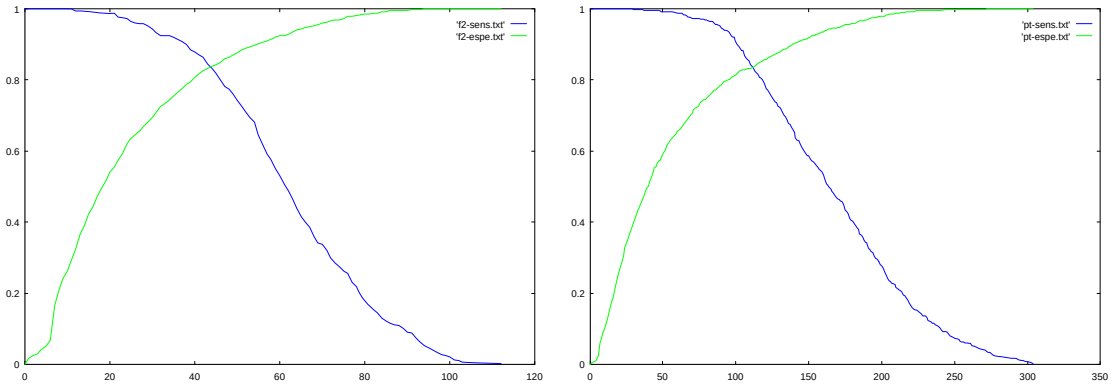

Figura 4.15. Relación entre sensibilidad y especificidad para factor 2 (izquierda) y puntuación total (derecha)

Como podemos observar la sensibilidad es una función decreciente, y la especificidad creciente, y hay un punto en el que se cruzan: en teoría, este punto correspondería a un punto óptimo (máximas sensibilidad y especificidad conjuntamente), con la matriz o los coeficientes de costes y beneficios que especifica por defecto BMDP. Como ya hemos visto, este concepto de optimización requiere considerar no sólo la razón costes-beneficios sino la prevalencia del resultado positivo en la población.

Pues bien, la curva ROC no es más que la representación de sensibilidad (eje de ordenadas) frente a 1-especificidad (eje de abscisa) para cada punto de corte. Utilizando la salida que produce el programa NCSS 2004 y la posibilidad de comparar en un mismo gráfico dos índices, tendremos el siguiente gráfico:

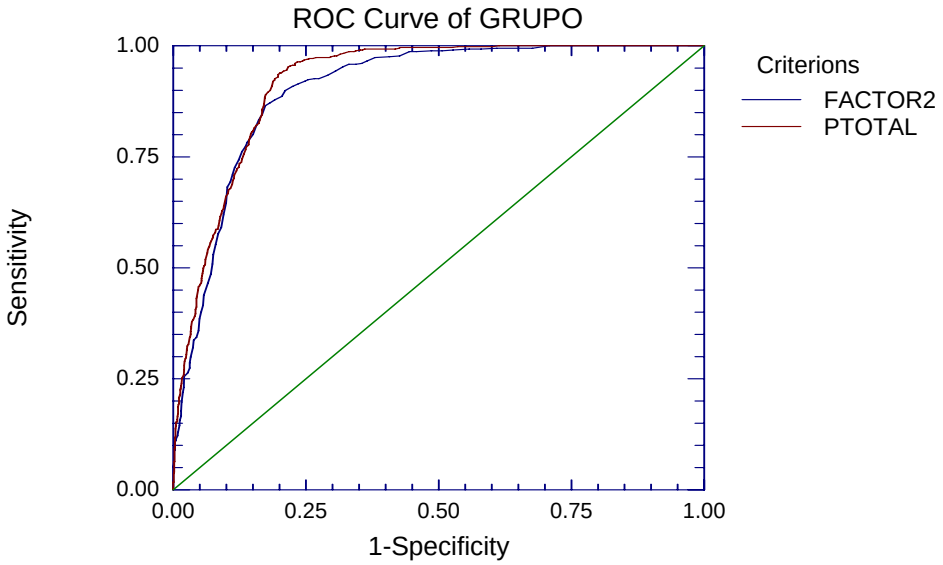

Figura 4.16. Curvas ROC empíricas para factor 2 y puntuación "suma" del instrumento de detección de maltrato infantil, obtenidas con el programa NCSS 2004

La curva ROC se puede interpretar como un gráfico de las tasas de verdaderos positivos frente a falsos positivos asociadas con **todos** los posibles puntos de corte o umbrales para clasificar un evento u objeto como positivo. Siempre se encuentra en el espacio cuadrado entre (0, 0) y (1, 1), al que algunas veces se denomina "espacio ROC", y tradicionalmente se dibuja también la diagonal positiva para indicar el punto de nula capacidad predictiva. Efectivamente, si la sensibilidad iguala a 1-especificidad a lo largo de todos los puntos de corte es que nuestra prueba simplemente selecciona unos casos de otros por puro azar.

Formalmente, no se trata de una curva como tal, sino de un conjunto de puntos (x, y) según la definición siguiente, para la cual introducimos la notación que usaremos posteriormente.

Sea x_{VP} la puntuación en la escala para un caso verdaderamente positivo, y sea x_{VN} la puntuación para un caso verdaderamente negativo. Asimismo, estableceremos valores de umbral o punto de corte, u_i, que irá variando en esa misma escala. La curva ROC empírica será el conjunto de puntos

$$\left(P\left[x_{VN} < u\right], P\left[x_{VP} \geq u\right]\right) \quad \forall u_i \tag{4.15}$$

o dicho de otro modo, cada punto (x, y) en el gráfico es

x = VN(u_i) proporción de verdaderos negativos, esto es, aquéllos que están por debajo del punto de corte en la prueba y realmente son negativos.

y = VP(u_i) proporción de verdaderos positivos, esto es, aquéllos que están igual o por encima del punto de corte y realmente son positivos.

Podemos establecer en este momento algunas conclusiones:

- La curva ROC resume en un único gráfico sensibilidad y especificidad **para todos los puntos de corte** en una prueba.
- Curva ROC en sí misma, “como conjunto de puntos”, es una medida única para una prueba, que resume la capacidad predictiva para la detección de casos

verdaderamente positivos. Se demuestra que el área bajo la curva ROC es una medida de la eficacia predictiva del modelo, independientemente del punto de corte que se establezca con la misma (Swets y Pickett, 1982; Swets, 1986).

- Mediante la curva ROC empírica (conjunto de puntos) podremos comparar visualmente el rendimiento de varias pruebas sobre los mismos casos, o sobre casos diferentes. Cuanto más alejada se encuentre la curva de la diagonal, mayor será la capacidad predictiva de la prueba.
- Asimismo, podremos decir si una prueba detecta más que por puro azar, simplemente comprobando que la curva se separa de la diagonal. En términos del área bajo la curva, nuestra curva debiera ser mayor de 0.5 que es el área de la diagonal, para tener alguna capacidad predictiva.

4.2.7 ¿Cómo se calcula y se interpreta el área bajo la curva ROC?

El área bajo la curva ROC, o AUC (siglas del nombre inglés "*Area Under Curve*"), es el estadístico que proporciona una medida completa de la capacidad predictiva de un instrumento o sistema diagnóstico. El valor que se obtiene tiene una interpretación directa, por ejemplo, un área de 0.84 significa que un individuo seleccionado aleatoriamente del grupo positivo "tiene un valor en la prueba mayor que el de un individuo elegido aleatoriamente del grupo negativo un 84% de las veces" (Zweig & Campbell, 1993).

Cuando el criterio de medida o indicador diagnóstico no puede distinguir entre los dos grupos o resultados, esto es, cuando no hay diferencia entre las dos distribuciones, el área será 0.5, y la curva ROC coincidirá con la diagonal positiva del "espacio ROC". Y en el otro caso extremo, cuando la capacidad predictiva sea perfecta, esto es, haya perfecta separación, ningún solapamiento entre las dos distribuciones, el área bajo la ROC es igual a 1, y alcanzará perfectamente la esquina superior izquierda del "espacio ROC".

Su cálculo se puede realizar fácilmente mediante la regla trapezoidal (Hanley y Mc. Neil, 1982), pero actualmente se realiza utilizando técnicas no paramétricas que además permiten su contraste estadístico.

4.2.8 Estimación del área bajo la curva: técnicas no paramétricas

Bamber (1975) encontró que, calculando el área bajo la curva ROC mediante la regla trapezoidal, el área que cae por debajo de los puntos que forman una curva ROC empírica es igual al estadístico U de Mann-Whitney para comparar distribuciones de valores de dos muestras. Por tanto, las áreas se pueden calcular con esta fórmula de Mann-Whitney, o con transformaciones adecuadas, con el estadístico de suma de rangos de Wilcoxon.

El siguiente trabajo sobre este particular (pruebas estadísticas no paramétricas para el área bajo la curva ROC) proviene de Hanley y McNeil (1982), sobre muestras independientes. Estos dos autores tienen también una publicación en 1983 en la que exponen el método para datos a partir de diseños con medidas relacionadas.

El primer trabajo de Hanley y McNeil muestra que si se realiza una prueba de Wilcoxon sobre las valoraciones (por ejemplo, las realizadas en un experimento de evaluación de un sistema de diagnóstico por la imagen), se obtiene la misma cantidad que si se calcula el área bajo la curva ROC utilizando la regla trapezoidal.

Definamos ahora

$$\theta = \Pr(x_{VP} > x_{VN}) \qquad (4.16)$$

o "verdadera área bajo la curva ROC".

El contraste estadístico de Wilcoxon, W, se utiliza normalmente para probar si los niveles de una variable x en una población (p.ej., VP) tienden a ser mayores que en una segunda población (p.ej., VN), sin tener que suponer la forma de la distribución de estas dos poblaciones.

La hipótesis nula es que x no es un discriminador útil, esto es, que un valor x de un individuo de la población VP es tan probable que sea tan pequeño como un valor x de un individuo de VN sea mayor que él, o dicho de otro modo que:

$$\theta = \Pr(x_{VP} > x_{VN}) = 0.5 \qquad (4.17)$$

Cuanto más cerca de 1, mejor discriminador es nuestro índice o escala x.

Sea $n_{VP(j)}$ la muestra de la población VP, y $n_{VN(k)}$ la muestra de la población VN. El procedimiento de Wilcoxon se basa en hacer todas las posibles comparaciones entre los valores $x_{VP(jk)}$ de $n_{VP(j)}$ * los valores $x_{VN(jk)}$ de $n_{VN(k)}$, de acuerdo con la regla:

$$S_{jk}(x_{VP}, x_{VN}) = \left\{ \begin{array}{c} 1 si\ x_{VP} > x_{VN} \\ 1/2\ si\ x_{VP} = x_{VN} \\ 0 si\ x_{VP} < x_{VN} \end{array} \right\} \qquad (4.18)$$

y promediando los sujetos o casos a lo largo de todas las comparaciones n_{VP} * n_{VN},

$$\theta = W = \frac{1}{n_{VP} \cdot n_{VN}} \sum_{j=1}^{n_{VP}} \sum_{k=1}^{n_{VN}} S_{jk}(x_{VP,} x_{VN}) \qquad (4.19)$$

Este cálculo no depende de los valores reales de x, sólo de sus ordenaciones.

Además de este estadístico que resume el área bajo la curva, nos interesa su error típico, dado que nuestro interés reside en cuantificar hasta qué punto es variable W, ó θ, en diferentes muestras de similar tamaño. Cuando $\theta > 0.5$, W deja de ser no paramétrico, su error típico o SE(W) depende de dos cantidades específicas de la distribución, Q_1 y Q_2:

Q_1= Probabilidad de que dos VP elegidas aleatoriamente sean ambas ordenadas como superiores que un caso VN elegido aleatoriamente.
Q_2= Probabilidad de que un caso VP elegido aleatoriamente sea ordenado como superior en el criterio que dos casos VN elegidos aleatoriamente.

Suponiendo, al igual que Green y Swets (1966) que las evaluaciones están en una escala que es suficientemente continua SE(W) se demuestra que es:

$$SE(W) = \sqrt{\frac{\theta(1-\theta) + (n_{VP} - 1)(Q_1 - \theta^2) + (n_{VN} - 1)(Q_2 - \theta^2)}{n_{VP} n_{VN}}} \qquad (4.20)$$

Es ésta la fórmula que usan programas que calculan estos estadísticos, como el NCSS (Hintze, 1998, 2003), aunque señalan que lo calculan mediante la "aproximación exponencial negativa a la fórmula exacta propuesta por Hanley (1982)".

Hanley y McNeil comparan 3 procedimientos en su artículo, con un ejemplo calculado en el mismo artículo:

- El área bajo la curva mediante la regla trapezoidal y mediante el estadístico W, que coinciden exactamente.
- El área bajo la curva suavizada obtenida mediante el algoritmo de máxima verosimilitud de Dorfman y Alf (1969), que sale superior por un pequeño margen (la ROC empírica suele subestimar la ROC suavizada).
- El área bajo la curva derivada de los parámetros de una línea recta ajustada a la ROC dibujada en papel probabilístico (los ejes son proporciones de la normal bajo la puntuación correspondiente), con lo que obtienen una cantidad intermedia.

Hanley y McNeil señalan que el hecho de que los estimadores W sean menores reflejan más que nada la naturaleza menos continua de las valoraciones en la escala de valoración propuesta en los estudios de diagnóstico por la imagen. En resumen, recomiendan el uso de este estadístico por su mayor simplicidad de cálculo (en aquella época los ordenadores no estaban tan extendidos como en la actualidad, lo que justifica también el ejemplo calculado a mano en su artículo), y porque la diferencia es suficientemente pequeña (menos de un 1%) con respecto a la versión suavizada de la ROC.

A este respecto señalan que en el rango de interés para los estudios ROC (con AUC's de 80% o más), el modelo exponencial negativo proporciona errores típicos que son ligeramente más conservadores que los otros modelos considerados. Bajo este supuesto, las cantidades Q_1 y Q_2 se pueden expresar mucho más simplemente como funciones de θ según:

$$Q_1 = \frac{\theta}{(2-\theta)}$$
$$Q_2 = \frac{2\theta^2}{(1+\theta)} \qquad (4.21)$$

Esta es la aproximación a la que se refiere Hintze (2003) y la que muy probablemente realiza el paquete estadístico NCSS.

Todas las fórmulas anteriores suponen datos estadísticamente independientes, esto es, no son adecuados para situaciones en las que dos "lectores" están examinando los mismos conjuntos de casos, o cuando un único "lector" está examinando el mismo conjunto de casos bajo distintos supuestos. Hanley y McNeil (1982) sugieren técnicas estadísticas para muestras emparejadas, al igual que sucede con las pruebas t de Student.

Aunque en 1983 Hanley y McNeil publican su segundo artículo en el que publican el método para el cálculo no paramétrico del área bajo la curva para muestras relacionadas, no lo recogeremos aquí porque preferimos mostrar el enfoque más general de DeLong *et al*. (1988), a continuación.

DeLong, DeLong y Clarke-Pearson publican en 1988 un método no paramétrico para comparar el área bajo dos o más curvas ROC correlacionadas (emparejadas). Se trata de uno de los artículos más citados en casi cualquier contexto que use curvas ROC, porque no sólo propusieron el método sino que publicaron una macro SAS para hacer los cálculos sobre la matriz de covarianzas y hacer contrastes, permitiendo así que numerosos investigadores pudieran usar el método directamente.

El método se ha hecho muy popular porque no establece ningún supuesto sobre la forma de la distribución, y su cálculo es diferente según sean muestras independientes o muestras relacionadas. Para el primer caso, el cálculo se realiza mediante el siguiente procedimiento (Hintze, 2003, p. 5100; recogiendo a McClish, 1989,y referenciando también Zhou *et al*., 2002). El punto de partida es la misma fórmula de puntuación que ya establecen Hanley y Mc. Neil, recogida aquí en la fórmula 4.18. Como en ocasiones

anteriores, definimos $x_{VP(jk)}$ la puntuación en la escala para un caso verdaderamente positivo, y sea $x_{VN(jk)}$ la puntuación en la escala para un caso verdaderamente negativo y en donde

j= 0,1 representa el grupo al que se le asigna o que se le predice y

k=1,2 son las 2 pruebas o escalas que se utilizan (el caso más sencillo).

Para muestras independientes se calculan entonces las varianzas de las estimaciones de área bajo la curva, siempre a partir del procedimiento establecido en 4.18 para calcular S_{jk} (adaptado a nuestra notación de Hintze, 2003):

$$V(x_{VP(jk)}) = \frac{1}{n_{VN} - 1} \sum_{j=1}^{n_{VN}} S_{jk}(x_{VP} x_{VN}),\ k = 1,2$$
$$V(x_{VN(jk)}) = \frac{1}{n_{VP} - 1} \sum_{k=1}^{n_{VP}} S_{jk}(x_{VP} x_{VN}),\ k = 1,2 \qquad (4.22)$$

El área para cada curva viene dada por:

$$A_k = \frac{\sum_{k=1}^{n_{VP}} V(x_{VN(jk)})}{n_{VP}} = \frac{\sum_{j=1}^{n_{VN}} V(x_{VP(jk)})}{n_{VN}},\ k = 1,2 \qquad (4.23)$$

Además tenemos:

$$S_{X(i)} = \frac{1}{n_{jk} - 1} \sum_{j=1}^{n_{jk}} \left[V(x_{X(i)}) - A_k\right]^2,\ i = VN, VP \quad j = 1, \quad k = 1,2 \qquad (4.24)$$

Y finalmente las varianzas para cada curva son:

$$V(A_k) = \frac{S_{XVP(k)}}{n_{VN(k)}} + \frac{S_{XVN(k)}}{n_{VN(k)}} \qquad (4.25)$$

En el caso de las muestras relacionadas, además de las varianzas necesitamos la covarianza entre las pruebas relacionadas, que viene dada por:

$$S_{XVP(1)XVP(2)} = \frac{1}{n_{VP} - 1} \sum_{j=1}^{n_{VP}} \left[V(x_{VP(1)}) - A_1\right]\left[V(x_{VP(2)}) - A_2\right]$$
$$S_{XVN(1)XVN(2)} = \frac{1}{n_{VN} - 1} \sum_{j=1}^{n_{VN}} \left[V(x_{VN(1)}) - A_1\right]\left[V(x_{VN(2)}) - A_2\right] \qquad (4.26)$$

Y a partir de ahí:

$$Cov(A_1, A_2) = \frac{S_{XVP(1)\,XVP(2)}}{n_{VP}} + \frac{S_{XVN(1)\,XVN(2)}}{n_{VN}} \qquad (4.27)$$

A continuación presentamos los resultados del análisis realizado por NCSS 2004, que presenta todos estos resultados, incluyendo intervalos de confianza y pruebas estadísticas de contraste (que veremos posteriormente) para los dos indicadores para los que estamos realizando el ejemplo: el factor 2 o predictor de negligencia y abandono y la puntuación total.

Tabla 4.7. Estimación de la AUC para los dos factores de detección de maltrato (salida del programa NCSS 2004)

Criterio	Estimador empírico de AUC	Error típico de AUC	Valor Z para probar que AUC >0.5	Prob. 1 cola	Prob. 2 colas	Porcentaje observado de casos positivos	N total
FACTOR2	0.90280	0.00720	55.96	<0.0001	<0.0001	0.30120	1743
PTOTAL	0.91787	0.00632	66.09	<0.0001	<0.0001	0.30120	1743

En la salida mostrada antes, el programa indica las siguientes precauciones, que es interesante señalar aquí:

- Este enfoque subestima la AUC cuando tenemos pocos valores del criterio (de 3 a 7).
- El valor de Z compara la AUC con el valor de predicción "inútil" de 0.5. Se utiliza la prueba de una sóla cola normalmente porque el interés del investigador será que el criterio rinda mejor que la diagonal, esto es, que él area sea sistemáticamente mayor que 0.5 (y no sólo "diferente").
- El contraste Z utilizado es preciso únicamente para muestras con al menos 30 sujetos en la condición positiva y otros 30 en la condición negativa.

Ambos indicadores son estadísticamente significativos, por lo que podemos concluir que predicen más que por simple azar.

A continuación también puede ser interesante mostrar los intervalos de confianza para las estimaciones de AUC empíricas:

Tabla 4.8. Intervalos de confianza (95%) de la AUC para los dos factores de detección de maltrato (salida del programa NCSS 2004)

Criterio	Estimador empírico de AUC	Error típico de AUC	Límite inferior (n.c. 95%)	Límite superior (n.c. 95%)	Porcentaje observado de casos positivos	N total
FACTOR2	0.90280	0.00720	0.88768	0.91597	0.30120	1743
PTOTAL	0.91787	0.00632	0.90455	0.92941	0.30120	1743

De los resultados anteriores, observamos que la puntuación total supera en capacidad predictiva al factor 2, aunque por un escaso margen. ¿Hay diferencias significativas entre ambas?

4.2.9 Contraste estadístico de curvas ROC empíricas

Al ser un índice completo de la capacidad diagnóstica de un instrumento, las curvas ROC permiten hacer comparaciones entre diferentes instrumentos en cuanto a su capacidad predictiva. Esto se puede hacer de forma visual, comparando las distintas curvas, pero resulta mucho más eficaz de forma numérica, comparando las áreas bajo la curva mediante un procedimiento de contraste estadístico tradiconal.

Con este objetivo se han propuesto enfoques que permiten hacer pruebas de significación estadística de la diferencia entre las áreas bajo dos curvas ROC, comenzando por el trabajo de Hanley y McNeil (1983) ya referenciado. Metz y Kronman (1980) recogen la historia de propuestas hasta aquel momento:

- Gourevitch y Galanter (1967) propusieron una prueba de significación estadística para comparación por pares de curvas ROC, basadas únicamente en el parámetro *d'* estimado de un punto operativo. Estas curvas ROC debían estar obtenidas a partir de distribuciones normales.

- Marascuilo (1970) extendió ese test para comparar una única curva ROC frente al azar, y para realizar múltiples comparaciones de 3 ó más curvas ROC, incluyendo

un procedimiento *post-hoc* apropiado. Una curva ROC que se pueda definir sólo por el parámetro *d'* es simétrica sobre la diagonal negativa del espacio ROC, e implica que las varianzas de las distribuciones gaussianas señal y señal+ruido son iguales.

Metz y Kronman (1980) señalan la poca aplicabilidad de estos enfoques, dado que los estrictos supuestos que establecen raramente se cumple en condiciones normales de las pruebas. Por tanto, la necesidad que es punto de partida de su trabajo son estas curvas ROC binormales, pero "asimétricas", que consideraremos en la tercera parte de este capítulo.

Los siguientes desarrollos son los de:

- Hanley y McNeil (1982, 1983) que desarrollan modelos no paramétricos para la comparación de las áreas bajo curvas ROC, obtenidas de diseños de grupos independientes, o de muestras relacionadas, respectivamente.
- DeLong *et al*. (1988) desarrollan completamente este último método, que recogen a su vez Zhou *et al*. (2002) y Hintze (2003), a partir de los cálculos de estimador del área y de su varianza expuestos anteriormente.

Este último enfoque es el que se utiliza más frecuentemente, y propone un contraste con la distribución normal tipificada (contraste Z), que asintóticamente sigue la distribución normal típica:

$$z = \frac{A_1 - A_2}{\sqrt{V(A_1 - A_2)}} \qquad (4.28)$$

donde

$$V(A_1 - A_2) = V(A_1) + V(A_2) - 2Cov(A_1, A_2) \qquad (4.29)$$

Para muestras independientes, las varianzas aparecen en la fórmula 4.25 más arriba, y la covarianza es 0. En el caso de pruebas con muestras relacionadas se calculará la covarianza según la fórmula 4.27 y se incluirá en la fórmula anterior. La interpretación de los contrastes es directa, como cualquier contraste con la normal tipificada.

La tabla siguiente muestra el resultado de este contraste. Podemos concluir que existen diferencias significativas entre ambos indicadores (z=3.06, p<0.01). La puntuación total ofrece un rendimiento diagnóstico significativamente superior.

Tabla 4.9: Contraste estadístico de igualdad de las estimaciones de las AUC obtenidas de forma no paramétrica

Criterio	Estimador empírico de AUC (1)	Estimador empírico de AUC (2)	Valor de la diferencia	Error Típico de la diferencia	Valor del E.C. Z	Prob.
FACTOR2, PTOTAL	0.90280	0.91787	-0.01507	0.00492	-3.06	0.0022
PTOTAL, FACTOR2	0.91787	0.90280	0.01507	0.00492	3.06	0.0022

El programa NCSS 2004 proporciona el contraste en los dos sentidos, quizá para facilitar la comprensión para personas que no estén acostumbradas a este tipo de contrastes. Lógicamente los resultados son los mismos.

4.2.8 Tamaño de la muestra para el análisis de curvas ROC empíricas

Hemos visto que los programas informáticos muestran notas de precaución sobre la estimación de los parámetros cuando tenemos pocos casos en nuestra base de datos.

Metz (1978) recomienda, en relación con el tamaño de la muestra para poder extraer conclusiones significativas del estudio, que la muestra debe ser de alrededor de 100 observaciones. Se debería requerir un mínimo de 50 casos en cada uno de los grupos, de tal modo que 1 caso represente no más del 2% de las observaciones de su grupo.

Hanley y McNeil (1982) intentan responder a la pregunta ¿cuántos casos se deben estudiar para asegurar un nivel aceptable de precisión en un estudio de evaluación de un sistema diagnóstico? O, en otras palabras, cuáles deben ser los tamaños de n_{VP} y n_{VN} para que el error típico resultante sea razonablemente pequeño, y por tanto el intervalo de confianza también razonablemente pequeño.

Siguiendo el razonamiento para la estimación de Q_1 y Q_2 mediante el modelo exponencial negativo, se pueden estimar los errores típicos para θ en función de los tamaños de muestra n_{VP} y n_{VN}. A partir de lo cual concluyen:

- Los errores típicos varían de manera inversamente proporcional a raíz cuadrada de n, de tal manera, que, por ejemplo, se debe cuadruplicar el tamaño de la muestra para reducir a la mitad el error típico.
- Los errores típicos son mínimos para θ muy altos, por ejemplo, cercanos a 1.
- Los errores típicos obtenidos con este procedimiento no paramétrico son ligeramente más conservadores que los obtenidos mediante el modelo gaussiano (o binormal).

¿Cómo se calcularía el error típico en un estudio 2AFC? En 2AFC, un estudio típico de TDS, se calcularía el área bajo la curva o θ calculando la fracción de los m pares de imágenes en los cuales los VP y VN casos fueron identificados correctamente, y se acompañaría este estimador con el error típico basado en la distribución binomial de

$$\sqrt{\hat{\theta}(1-\hat{\theta})/m} \qquad (4.30)$$

Para conseguir un error típico de 0.03, y suponiendo que el estimador del área bajo la curva fuera .9 o 90% en un estudio 2AFC, harían falta m=100 parejas de casos, un número mucho mayor, señalan Hanley y McNeil, que el que se utilizó en su estudio, 109.

Otra pregunta de interés es ¿Cuántos sujetos son necesarios para comparar las áreas bajo dos curvas ROC? Finalmente, Hanley y McNeil proporcionan una útil tabla de tamaños de muestra necesarios para obtener ciertos niveles "tipo" de áreas bajo la curva (entre 0.700 y 0.950), para potencias de 80, 90 ó 95% de detectar diferencias entre dos curvas (Hanley y McNeil, 1982, p. 34).

Las tablas se obtuvieron utilizando la siguiente fórmula:

$$n = \left[\frac{Z_\alpha \sqrt{2V_1} + Z_\beta \sqrt{V_1 + V_2}}{\delta} \right]^2 \qquad (4.31)$$

donde

$Z_\alpha = 1.645$, para una prueba unilateral de significación al 5%

$Z_\beta = 0.84, 1.28, o 1.645$, para potencias de 80%, 90% o 95% respectivamente

$\delta = \theta_2 - \theta_1$

$V_1 = Q_1 + Q_2 - 2\theta_1^2$

$V_2 = Q_1 + Q_2 - 2\theta_2^2$, (en ambos casos según los cálculos de Q1 y Q2 según modelo exponencial negativo anteriormente expuesto)

4.3 Curvas ROC binormales

Las curvas ROC que observamos tradicionalmente en la literatura de la TDS son mucho más "suaves" que las que hemos visto en el ejemplo anterior. De hecho, realmente son curvas y no un conjunto de puntos que se aproxima a una curva. Esto es porque en la TDS como ya hemos visto, se suponía que las dos distribuciones, la de señal y la de señal + ruido, son normales y con la misma varianza. En ese caso, la curva ROC es simétrica sobre la diagonal negativa del espacio ROC y se puede demostrar que la curva ROC viene definida por un único parámetro, *d'*, siempre con el supuesto de distribuciones normales y homoscedásticas de partida.

Metz y Kronman (1980) señalan que que este supuesto raramente se cumple en condiciones normales de las pruebas en Medicina. Por tanto, la necesidad que es punto de partida de su trabajo son estas curvas ROC binormales, pero “asimétricas”.

El método propuesto originalmente por Metz (1978) considera dos poblaciones, una de verdaderos positivos y otra de verdaderos negativos. Se asume que la variable criterio, o una transformación de la misma, siguen una distribución normal en cada población. Muchas investigaciones han mostrado a través de estudios de simulación que este supuesto de "binormalidad" no es limitador como puede parecer a simple vista puesto que datos con distribuciones no normales pueden a menudo ser transformados de tal modo que muestren una distribución casi normal.

Un modelo para una curva ROC binormal, asimétrica, debe incluir dos parámetros (frente al único, *d’*, de las curvas ROC binormales simétricas), uno para la razón de las

desviaciones típicas de las distribuciones de partida (VP, o señal y VN, sólo ruido), expresadas en unidades de una u otra.

Una exposición apropiada para este contexto es la de Metz y Pan (1999). Por otro lado, tendremos las fórmulas recogidas en Hintze (2003), y tomadas a su vez de McClish (1989). Ambas tratan de lo mismo y son complementarias al menos en cuanto a su notación:

Supongamos dos poblaciones, una de individuos positivos en el evento que queremos detectar y otra de individuos negativos. Supongamos además que el valor de la variable criterio está disponible para todos los individuos en la población. Sea x|VN el valor de la variable criterio para la población negativa e y|VP el valor de la variable criterio en la población positiva. El modelo binormal supone que X e Y se distribuyen según la normal con diferentes medias y varianzas (lo que Metz llama un "modelo binormal impropio", frente al de igual varianzas que supone la TDS clásica). Tendremos:

$$f(x \mid VN) = \frac{1}{\sqrt{2\pi}} \left\{ \exp - \frac{x^2}{2} \right\} \tag{4.32}$$

para los casos verdaderos negativos, y para los verdaderos positivos:

$$f(y \mid VP) = \frac{1}{\sqrt{2\pi}} \left\{ \exp - \frac{(bx - a)^2}{2} \right\} \tag{4.33}$$

Donde:

$$a = \frac{\mu_{VP} - \mu_{FP}}{\sigma_{VP}} \tag{4.34}$$

$$b = \frac{\sigma_{FP}}{\sigma_{VP}} \tag{4.35}$$

Y por tanto, tenemos las dos distribuciones de puntuaciones (Hintze, 2003):

$$\begin{aligned} X &\sim N(\mu_{FP}, \sigma_{FP}) \\ Y &\sim N(\mu_{VP}, \sigma_{VP}) \end{aligned} \tag{4.36}$$

La curva ROC biene definida por la función (Hintze, 2003):

$$\{FP(c), VP(c)\} = \left\{ \phi\left(\frac{\mu_{FP} - c}{\sigma_{FP}} \right), \phi\left(\frac{\mu_{VP} - c}{\sigma_{VP}} \right) \right\}, \quad -\infty < C < \infty \tag{4.37}$$

donde $\phi(z)$ es la función de distribución normal acumulada.

El área de la curva ROC se puede demostrar a partir de las fórmulas anteriores (Metz y Pan, 1999, Hintze, 2003) que es:

$$A_z = \phi\left(\frac{a}{\sqrt{(1+b^2)}} \right) \tag{4.38}$$

Muy importante, porque puede ser muy interesante para aplicaciones prácticas, el área bajo una porción de la curva, entre los puntos de corte c_1 y c_2, viene dada por:

$$A = \int_{c_1}^{c_2} VP(c) FP'(c)\, dc = \frac{1}{\sigma_{FP}} \int_{c_2}^{c_1} \left(\phi\left(\frac{\mu_{VP} - c}{\sigma_{VP}} \right) \phi\left(\frac{\mu_{FP} - c}{\sigma_{FP}} \right) \right)\, dc \tag{4.39}$$

Hintze (2003) señala que el área parcial bajo una curva ROC se define normalmente en términos de un rango de tasas de falsos positivos más que en límites de criterio c_1 y c_2. Sin embargo, existe una relación uno-a-uno entre estas dos cantidades, que viene dada por:

$$c_i = \mu_{FP} + \sigma_{FP} \phi^{-1}(FP_I) \tag{4.40}$$

y que permite que se calculen límites del criterio de las tasas deseadas de falsos positivos. El programa NCSS 2004 ofrece una caja de diálogo para introducir estos puntos entre los cuales se calculará el área parcial bajo una curva ROC.

La varianza de A se deriva utilizando el método de diferenciales según expone con detalle Hintze (2003), recogiendo a su vez Zhou *et al*. (2002), y que no detallaremos aquí por su excesiva complejidad.

Una vez tenemos los estimadores del área bajo la curva ROC y de su varianza, Â y V(Â), se pueden calcular los intervalos de confianza y las pruebas de hipótesis estadísticas utilizando los métodos tradicionales. Sin embargo, Zhou *et al*. (2002) señalan la siguiente transformación que produce resultados en los estadísticos calculados que son más próximos a la normalidad, y asegura que los límites de los intervalos de confianza están fuera del rango cero-uno.

Las fórmulas, según las recoge Hintze (2003) son:

$$\hat{\psi} = \frac{1}{2}\ln\left(\frac{1+A}{1-A}\right) \qquad (4.41)$$

La varianza de este estimador se puede estimar mediante:

$$V(\hat{\psi}) = \frac{4}{\left(1-\hat{A}^2\right)^2} V(\hat{A}) \qquad (4.42)$$

Y se puede construir entonces un intervalo de confianza 100(1-α)% mediante:

$$L,U = \hat{\psi} \pm z_{1-\alpha/2}\sqrt{V(\hat{\psi})} \qquad (4.43)$$

4.3.1 Contraste de hipótesis con curvas ROC binormales

Cuando suponemos un modelo de curvas ROC binormales, los contrastes de hipótesis son muy sencillos. Si se trata de medidas independientes, la covarianza es 0, y por tanto la varianza de la diferencia en términos de AUC es la suma de las varianzas:

$$V(A_1 - A_2) = V(A_1) + V(A_2) \qquad (4.44)$$

Cuando se utiliza un diseño de muestras pareadas y por tanto correlacionadas, la fórmula es la misma que en el caso de ROC empíricas (véase la fórmula 4.29), pero la covarianza requiere unas fórmulas específicas. Hintze (2003) recoge los resultados de Mc Clish (1989), quien utilizó simulaciones para estudiar la precisión de la aproximación a la normal de los estadísticos z tal y como normalmente se proponen para hacer contrastes relacionados con el AUC de una curva, y que encontró que una transformación logística resultaba en un estadístico z más cercano a la normalidad.

La transformación propuesta es:

$$\theta(A) = \ln\left(\frac{FP_2 - FP_1 + A}{FP_2 - FP_1 - A}\right) \tag{4.45}$$

cuya versión inversa es:

$$A = (FP_2 - FP_1)\frac{e^{\theta} - 1}{e^{\theta} + 1} \tag{4.46}$$

Entonces la varianza estimada viene dada por:

$$V(\theta) = \left(\frac{2(FP_2 - FP_1)}{(FP_2 - FP_1)^2 - A^2}\right)^2 V(A) \tag{4.47}$$

Y cuya covarianza es:

$$\operatorname{cov}(\theta_1, \theta_2) = \left(\frac{4(FP_2 - FP_1)^2}{\left((FP_2 - FP_1)^2 - A_1^2\right)\left((FP_2 - FP_1)^2 - A_2^2\right)}\right) \operatorname{cov}(A_1, A_2) \tag{4.48}$$

El estadístico z ajustado es entonces:

$$z = \frac{\theta_1 - \theta_2}{\sqrt{V(\theta_1 - \theta_2)}} = \frac{\theta_1 - \theta_2}{\sqrt{V(\theta_1) + V(\theta_2) - 2\operatorname{cov}(\theta_1, \theta_2)}} \tag{4.49}$$

4.3.2 Ejemplo de representación de curvas ROC binormales para la detección de redención de puntos en una tarjeta de fidelización

El objetivo de esta tesis es realizar un análisis empírico de la predicción de redención de una tarjeta de fidelización. Este análisis se lleva a cabo en el capítulo 6, pero resulta interesante hablar aquí de la representación de la curva ROC binormal, para lo que, aun cuando omitimos los razonamientos que nos llevan a elegir un indicador u otro, presentamos los resultados en este punto.

La curva ROC binormal se presenta en la figura 4.17. superpuesta a la curva ROC empírica. Podemos observar las sutiles diferencias entre ambas, en especial el hecho de que la curva binormal suele proporcionar mayor área bajo la curva. En este caso no está

tan claro, pero los resultados en la literatura sugieren que la curva empírica subestima el área bajo la curva.

Un comentario sobre el dibujo de curvas ROC. Tradicionalmente se dibuja la diagonal positiva (y=x), pero además Swets suele dibujar una parte de la diagonal negativa. ¿Cuál es el motivo? Si la curva es simétrica con respecto a esta diagonal negativa en el gráfico, las distribuciones subyacentes tienen igual varianza. Según se desvíe la proporción de varianzas de 1, mayor será la asimetría. Esta diagonal negativa sirve para poder comprobar este hecho.

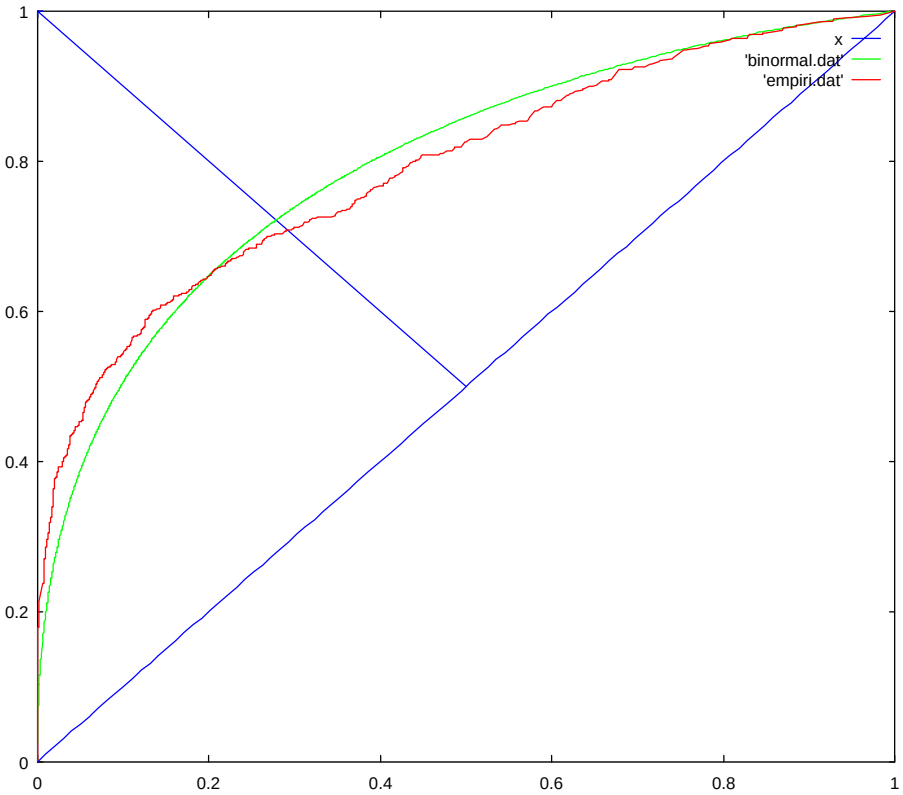

Figura 4.17. Curvas ROC empírica y binormal para el indicador de redención de puntos Ln(total de puntos conseguidos)

Otra forma de representar la curva ROC es en el espacio definido por las puntuaciones típicas de la distribución normal, esto es, transformaremos los valores de sensibilidad y (1-especificidad) a su equivalente en el área por debajo de la distribución normal tipificada. Esta representación produce, en el caso de curvas ROC binormales, una recta, y en el caso de las curvas ROC empíricas, una colección de puntos que, en función de su aproximación a distribuciones normales subyacentes, se acercará más o menos a la recta de la ROC binormal. La figura 4.18 a continuación presenta esta representación

para el mismo indicador que tratamos antes. Es ésta la famosa representación en "papel binormal" a la que se refiere Swets a menudo en sus publicaciones.

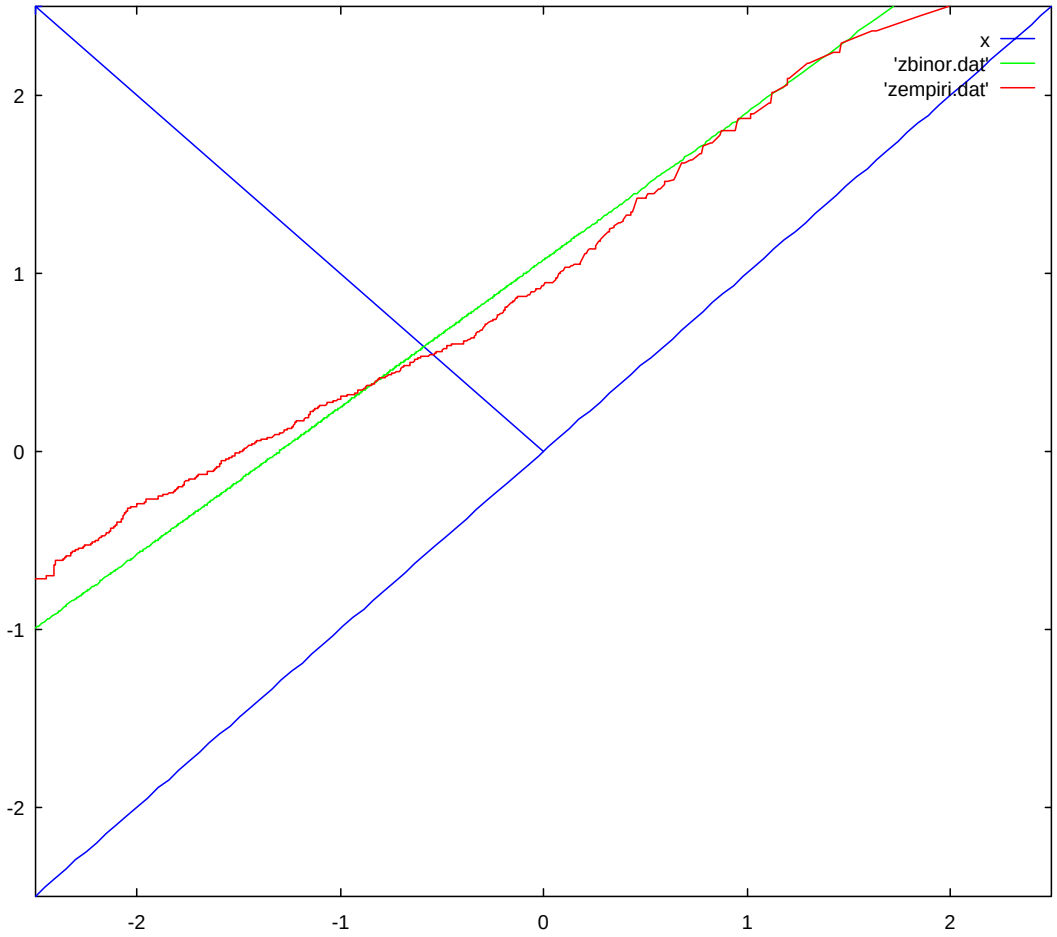

Figura 4.18. "Rectas" ROC empírica y binormal para el indicador de redención de puntos Ln(total de puntos conseguidos) en el espacio definido por las puntuaciones típicas correspondientes a los valores de sens. y (1-espec.)

Por supuesto, cuanto más alejadas estén nuestras distribuciones subyacentes de la distribución normal, mayor será la desviación con respecto a una recta. Para ilustrar este hecho, mostraremos en la figura 4.19 esta representación para los datos del factor 2 del instrumento de detección de maltrato que hemos puesto como ejemplo para la estimación de ROC empíricas.

Un análisis muy detallado de las distintas formas de distribuciones que subyacen a las diferentes formas ROC, en sus representaciones sobre escalas convencionales de

sensibilidad y especificidad, y en puntuaciones típicas se puede encontrar en Swets (1986).

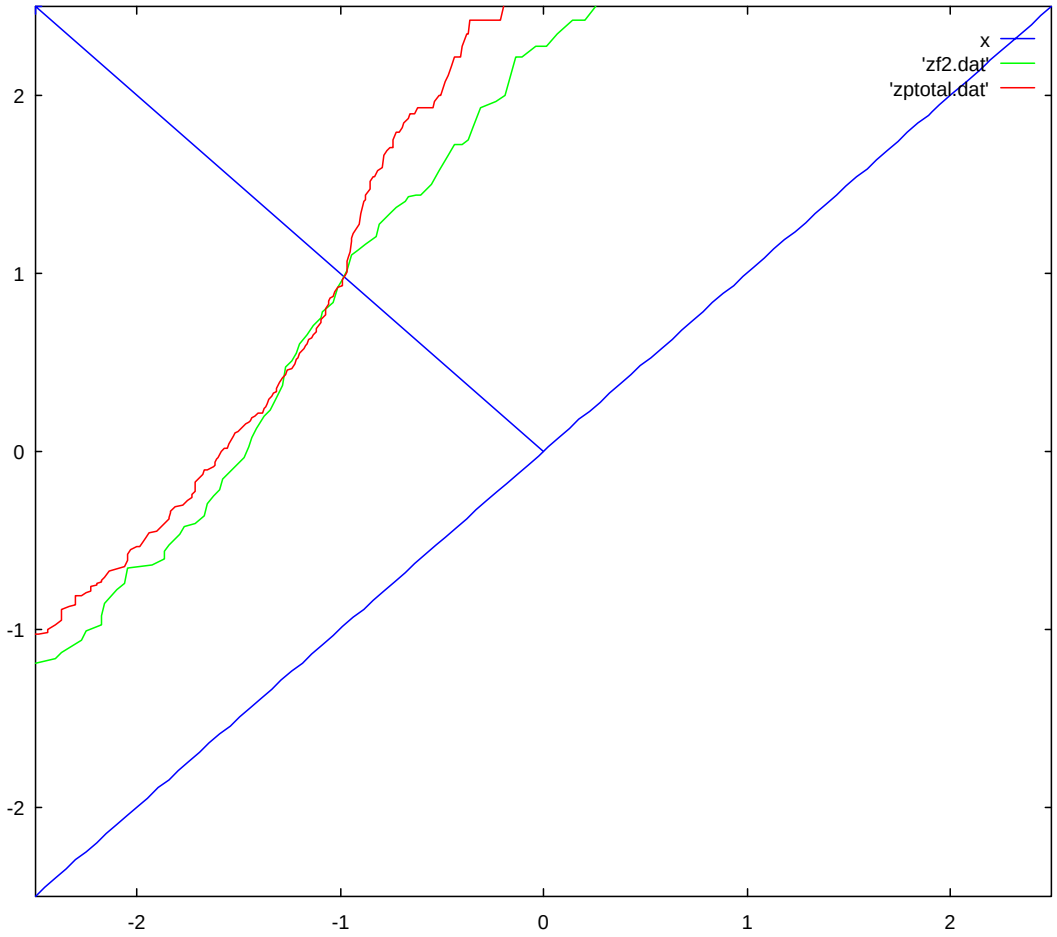

Figura 4.19. Representación de la ROC empírica para los datos del factor 2 y de la puntuación total del instrumento de detección de maltrato en el espacio definido por las puntuaciones típicas correspondientes a los valores de sens. y (1-espec.)

4.4 Comparación entre los procedimientos binormal y no paramétricos para el análisis de curvas ROC

En la práctica, las curvas ROC pueden ser obtenidas mediante los dos procedimientos que hemos visto, a los que se refieren normalmente como paramétricos (o binormal) y no paramétricos (o curva ROC empírica). Los primeros son expuestos principalmente en las obras de Swets y colaboradores (Swets y Pickett, 1982) y posteriormente de Metz y otros; mientras que las segundas se han desarrollado en los campos aplicados del diagnóstico clínico (Hanley y Mc. Neil, 1982, 1983) y su simplicidad de aplicación y generalidad han hecho de estas técnicas algo frecuente en el diagnóstico médico. Su aplicación al diagnóstico psicológico es asimismo cada vez más frecuente, especialmente en campos de investigación en Psicología más próximos al médico (Rice y Harris, 1995; Navarro y Doménech, 1996).

Estos últimos autores describen muy bien las ventajas y desventajas de los dos enfoques para el cálculo de las curvas ROC:

Tabla 4.10: Ventajas y desventajas de los enfoques paramétrico y no paramétrico para la estimación de curvas ROC (tomado de Navarro y Doménech, 1996).

Método y sus propiedades	**Ventajas**	**Desventajas**
Paramétrico (Swets y Picket, 1982): - Supuesto de distribución normal para los dos grupos (control / riesgo). - Parámetros que se modelizan son diferencia de medias y razón de varianzas. - Se obtienen curvas ROC mediante estimación de máxima verosimilitud. - Tradicionalmente se ha partido de datos cualitativos.	- Produce efectivamente una curva. - Compara curvas para cualquier sensibilidad y especificidad.	- La necesidad de cumplir supuestos exigentes. - Cálculo complejo. - Puede dar estimaciones sesgadas.

Tabla 4.10 (cont.): Ventajas y desventajas de los enfoques paramétrico y no paramétrico para la estimación de curvas ROC (tomado de Navarro y Doménech, 1996).

Método y sus propiedades	**Ventajas**	**Desventajas**
No paramétrico (Hanley y Mc. Neil, 1982, 1983): - Tradicionalmente el punto de partida ha sido un indicador cuantitativo continuo. - A partir de la representación gráfica de la sensibilidad y especificidad asociadas a cada punto de corte dentro de la zona de solapamiento de las puntuaciones del test en los dos grupos. - El cálculo del área bajo la curva ROC se realiza mediante la regla trapezoidal, y se demuestra que comparte las propiedades del estadístico de Wilcoxon (Hanley y Mc. Neil, 1982), o del estadístico U de Mann-Whitney para distribuciones de valores de 2 muestras (DeLong, DeLong y Clarke-Pearson, 1988).	- Usa toda la información . - No hace supuestos sobre las distribuciones de las puntuaciones en los grupos. - El cálculo, y la comparación de resultados para diferentes curvas, son mucho más sencillos.	- No proporciona efectivamente una curva, sino un conjunto de puntos. - Sistemáticamente, subestima el área bajo la curva ROC, el principal indicador de eficacia diagnóstica (DeLong, DeLong y Clarke-Pearson, 1988). - Compara curvas ROC sólo para sensibilidades y especificidades observadas.

4.5 Avances en la metodología de curvas ROC

Desde un primer momento desde su aparición en la literatura de diagnóstico, el interés de los investigadores era desarrollar procedimientos estadísticos no paramétricos para el análisis de curvas ROC (o de estudios de detección).

Resulta interesante el artículo de Morgan (1976), quien investiga la robustez de la TDS con respecto a la forma de las distribuciones subyacente. Normalmente estas distribuciones se suponen normales, y aquí se examina un modelo TDS con dos distribuciones uniformes que se solapan, en el caso de un experimento sí/no y un método experimental de valoración en una escala. Posteriormente, y ya en el ámbito de la Medicina, Long y Waag (1981) resumen las limitaciones que se han mostrado contra el uso del modelo conceptual y las medidas de rendimiento de TDS en varias situaciones de investigación aplicadas en Medicina. Se discuten los supuestos estadísticos que subyacen al cálculo de d' y β como estimadores independientes de la sensbilidad y del criterio, y se cuestiona la aplicabilidad total del modelo TDS en sí mismo. Se presentan resultados de 3 experimentos (vigilancia, búsqueda visual y detección auditiva) para ilustrar las consecuencias de una mala aplicación de TDS.

Desde un primer momento, resultan cruciales los avances en el desarrollo de programas informáticos que puedan estimar las curvas ROC. Además del mencionado de Dorfman y Altman (1969), destacaremos aquí el artículo Dorfman y Berbaum (1986), en el que describen el programa RSCORE J. Este programa estima estimadores *jackknife* de parámetros ROC y sus errores típicos sobre un grupo de observadores.

El análisis de curvas ROC ha provocado relativamente poco debate metodológico. Más bien podríamos decir que ha sido relativamente ignorado, especialmente en aquellos ámbitos con mayor enfoque hacia la "estadística tradicional". Uno de los pocos artículos críticos es el de Nelson (1986), quien intenta clarificar las diferencias entre los enfoques propio y de Swets a la evaluación de medidas de precisión discriminatoria. El enfoque de Swets se mantiene sobre el supuesto de que la predicción discriminativa permanece constante a lo largo de todos los umbrales de decisión (i.e, a lo largo de la ROC) y Nelson cuestiona la validez empírica de este supuesto en Psicología. Porque Swets usó este supuesto mientras que generaba ROCs teóricas para varias medidas, las discrepancias entre esas ROCs y las empíricas se puede deber a la falta de constancia psicológica que suponía.

Claramente, la década de los 80 supone el asentamiento definitivo en la Medicina, gracias en gran medida a los artículos clásicos de Hanley y Mc. Neil (1982, 1983) y

DeClarke *et al*. (1989), y en gran medida por la proposición de métodos de fácil cálculo, aplicables en un rango amplio de problemas de invetigación.

En los años 90 comienzan a aparecer estudios teóricos o metodológicos de aplicación de curvas ROC para evaluación del output de algoritmos matemáticos, como las redes neuronales: Eijkman (1992), por ejemplo, propone una medida de la sensibilidad de una red neuronal. Utiliza las medidas de la TDS aplicadas a una red de tres capas con error propagado hacia atrás, y corrección de pesos utilizando patrones de letras distorsionados por el ruido. Se midió la sensibilidad de la red en todas las etapas durante el proceso de aprendizaje.

Muchos autores han destacado las dificultades para encontrar información de buena calidad, sobre todo didáctica y aplicable, y han echado especialmente de menos un libro sobre esta técnica. Pues bien, no será hasta 2002 que se publique ese libro (Zhou, Obuchowski, y McClish (2002): Statistical Methods in Diagnostic Medicine). A este libro se refiere Hintze (2003) como básico para el desarrollo del software NCSS 2004. En una revisión de S. M. Rudolfer para Biometrics, se destaca asimismo que es el primer libro que intenta resumir los resultados conocidos en este campo.

Durante la década en que nos encontramos se produce un enorme incremento en el uso de las curvas ROC en lo que se ha venido en llamar "minería de datos". Destacamos los siguientes artículos que pueden ser interesantes:

- **Eguchi y Copas (2002):** Exponen que en análisis discriminante el lema de Neyman-Pearson establece que la curva ROC para una función lineal arbitraria es siempre por debajo de la curva ROC para la verdadera razón de verosimilitud. El área ponderada entre estas dos curvas se puede usar como una función de riesgo para encontrar buenas funciones discriminantes. La función de ponderación corresponde al objetivo del análisis, por ejemplo, minimizas el coste esperado de la mala clasificación, o maximizar el área bajo la ROC. Las funciones discriminantes resultantes pueden estimarse mediante regresión logística iterativamente reponderada.

- **Thompson (2003) c**onsidera la evaluación de la precisión diagnóstica completa de una secuencia de tests. La complejidad aumenta cuando se utilizan dos o más tests continuos, y se presentan varios enfoques para reducir la dimensionalidad. Una alternativa posible es ajustar el umbral para sucesivas visitas de acuerdo a características individuales. Estas posibilidades representan una porción particular de la superficie ROC, correspondientes a todas las posibles combinaciones de umbrales del test. Se concentran en el desarrollo y ejemplos de la situación en la que una prueba global se define como positiva si uno cualquiera de los individuos es positivo.

- **Dodd y Pepe (2003b)** destacan que, en Medicina, las nuevas pruebas diagnósticas o de *screening* se deben evaluar en función de su capacidad para discriminar los estados enfermos de no enfermo. El área parcial bajo la curva ROC es una medida de precisión de la prueba diagnóstica. Se presenta en este artículo una interpretación del área parcial bajo la curva, que da lugar a un estimador no paramétrico. Este estimador es más robusto que los estimadores existentes, que establecen supuestos paramétricos. Se muestra que se gana en robustez con sólo una moderada pérdida en eficiencia. Se describe un esquema de modelos de regresión para hacer inferencia sobre los efectos de las covariantes en el área bajo la curva parcial (AUC). Tales modelos pueden refinar el conocimiento sobre la precisión de la prueba. Los parámetros del modelo se pueden estimar utilizando métodos de regresión binaria. Se utiliza el esquema de regresión para comparar dos marcadores biológicos de antígenos específicos de próstata y para evaluar la dependencia de la precisión en el tiempo previo a la diagnosis clínica del cáncer de próstata.

- **Qu, Bao-ling, *et al*. (2003)** presentan un método de reducción de datos utlizando una "*wavelet transform*" en análisis discriminante cuando el número de variables que se analizan es mucho mayor que el número de observaciones, y se ilustra con un estudio de cáncer de próstata. En este estudio el tamaño de la muestra es de 248, pero el número de variables es 48538. El método identificó 11 de los 1271 "*wavelet coefficients*" con el mayor poder discriminatorio, y se probó mediante área bajo curva ROC (sensibilidad del 97% y especificidad del 100%). Revisan PCA (Análisis de Componentes Principales), *Discrete Wavelet Transform* (DWT) y Análisis Discriminante.

Resulta enormemente interesante la aplicación que surge para realizar meta-análisis de estudios de precisión de pruebas diagnósticas:

- **Dukic y Gatsonis (2003)** señalan en el resumen que los métodos meta-analíticos actuales para evaluar la precisión de las pruebas diagnósticas son generalmente aplicables a una selección de estudios que informan sólo de los estimadores de sensibilidad y especificidad, o a lo sumo, a estudios cuyos resultados se muestran utilizando un número igual de categorías ordenadas. En este artículo se propone un nuevo método meta-analítico para evaluar la precisión de una prueba y llegar a una curva ROC resumen para una colección de estudios que evalúen pruebas diagnósticas, incluso cuando los resultados de la prueba se informen en un número desigual de categorías ordenadas no anidadas. Se discuten formulaciones bayesianas y no bayesianas del enfoque. En el bayesiano se proponen varias formas de construir curvas ROC resumen y sus bandas creíbles. Se ilustra el enfoque con datos de un meta-análisis de progesterona para diagnóstico de embarazo. Se trata de un artículo muy interesante que además de recoger muy bien las bases de ROC (incluye referencia a Swets y Picket, 1982 -algo muy poco común-), habla del método delta y compara muchos estudios diferentes, cuya eficacia aparece representada como en la figura 4.20.

Por otro lado, aparecen aplicaciones muy interesantes sobre precisión en estudios longitudinales:

- **Rutter y Miglioretti (2003)** quienes presentan un método para estimar la precisión de las escalas de detección temprana utilizando curvas ROC y estadísticos asociados. Estas escalas de detección son típicamente semi-continuas con un rango conocido, con distribuciones casi simétricas cuando la condición de target está presente, y altamente sesgadas cuando la condición está ausente. Se modelizan los resultados de la escala de detección utilizando distribuciones normales truncadas para acomodar a estas diferentes formas de distribuciones y utilizando efectos aleatorios específicos para ajustar múltiples evaluaciones intra-individuos.

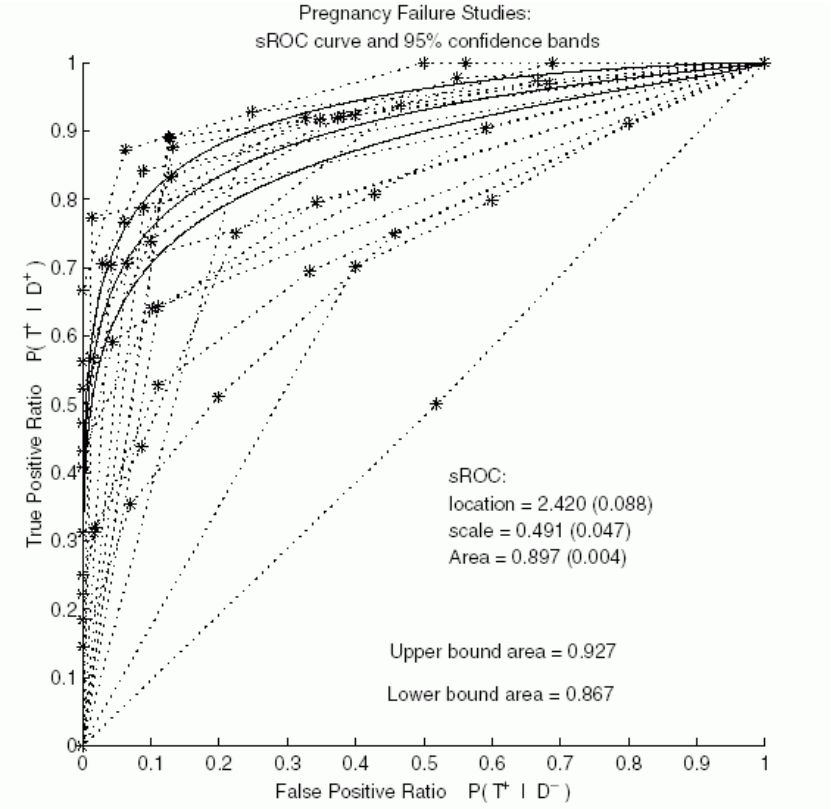

Figura 4.20. Meta-análisis de estudios de fallos en detección de embarazo. Tomado de Dukic y Gatsonis (2003)

Resultan muy interesantes los desarrollos sobre la aplicación conjunta de curvas ROC y regresión logística:

- Qin y Zhang (2003). En este artículo, Qin y Zhang exploran un modelo semiparamétrico asumiendo un modelo de razón de la densidad para las densidades de casos de enfermos frente a no-enfermos. Este modelo tiene una conexión natural con el modelo de regresión logística. El enfoque semiparamétrico propuesto es más robusto que un enfoque completamente paramétrico y es más eficiente que un enfoque completamente no paramétrico. Se demuestra mediante dos ejemplos reales que la curva ROC estimada por el modelo semiparamétrico es mucho más suave que la estimada por el método no paramétrico.

Además, sorprende la comparación establecida entre los enfoques de ROC y los de Youden y Taguchi para la evaluación de la calidad de las pruebas diagnósticas, por Tolga y Jiju (2000). Tolga y Jiju llevaron a cabo una comparación de las razones señal-

ruido desarrolladas por Taguchi en la ingeniería de la calidad, y el rendimiento de sistemas en la industria manufacturera. Se calculó un híbrido y se propuso su relevancia para los médicos como un método eficiente de evaluación.

En los últimos años han comenzado a aparecer aplicaciones de riesgo, tanto en ámbitos financiero como otros (seguros, etc.) además de por supuesto el médico:

Zhu1, Beling y Overstreet (2002) exploran un enfoque bayesiano para construir combinaciones de las salidas de clasificadores, como procedimiento para mejorar los resultados globales de clasificación. Proponen el esquema bayesiano para estimar la probabilidad posterior de estar en una cierta clase dados múltiples clasificadores. Este enfoque, que emplea modelización meta-gaussiana, pero no hace supuestos sobre la distribución de las salidas de los clasificadores, les permite capturar las dependencias no lineales entre los clasificadores e individuos combinados. Una propiedad importante del método propuesto es que produce un clasificador combinado que domina los individuos sobre los que se basan, en términos de riesgo bayesiano, tasa de error, y curva ROC. Para ilustrar este método, se muestran resultados empíricos de la combinación de las puntuaciones de crédito generadas a partir de varios modelos de *scoring*.

4.6 Conclusiones

Las curvas ROC, frente a otras técnicas con objetivos similares, presentan las siguientes ventajas, expuestas en Swets (1982):

1. Proporcionan un índice de eficacia diagnóstica puro, cualquiera que sea el criterio o punto de corte en el indicador cuantitativo en que se basa la decisión, e incluso independientemente de que dicho indicador esté sesgado. Esta medida resulta extremadamente útil para poder elegir una técnica entre varias en competencia.
2. Estima la probabilidad de diferentes resultados en la tabla de 4 resultados posibles. De este modo, permiten aislar causas de error en sistemas diagnósticos (por ejemplo, falsas alarmas y omisiones).
3. Proporciona un indicador de criterio de decisión (punto de corte o umbral para la toma de decisión), que permite incluir probabilidades (incluso estimaciones

subjetivas de la probabilidad) y costos o utilidades. Este método permite establecer razonadamente reglas o mecanismos óptimos de toma de decisión dentro de un sistema de diagnóstico. El área bajo la curva no depende de la elección del punto de corte o umbral.

Para nuestros objetivos, mostraremos cómo:

- Las curvas ROC, meramente descriptivas, son un magnífico instrumento para comparar visualmente diferentes índices, facilitando enormemente la tarea de decidir qué componentes de una escala compuesta serán más eficaces para nuestra tarea de predicción.
- El enfoque o modelo de curvas ROC no paramétrico es muy aplicable para muchos conjuntos de datos de la vida real, puesto que no requiere el cumplimiento de supuestos distribucionales.
- El enfoque paramétrico de análisis de curvas ROC es generalmente más potente, aunque más complejo también pues exige distribuciones normales de partida.
- El análisis de curvas ROC es quizá el mejor enfoque para estudiar la eficacia predictiva de un sistema diagnóstico, y además permite integrar el análisis coste-beneficio perfectamente.

A continuación revisaremos los diferentes índices para medir la capacidad diagnóstica, que nos reforzará en nuestra idea de las curvas ROC como mejor enfoque para esta tarea, para a continuación entrar en la parte analítica de esta tesis.