Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matematicas

Metodología para la evaluación del poder descriptivo de un MARCADOR mediante el análisis ROC (Receiver-Operating Characteristic)

T E S I S

para obtener el título de:

Licenciatura en Matemáticas Aplicadas

P R E S E N T A :

Oscar Méndez Cuanalo

directores de tesis:

Dra. Hortensia Josefina Reyes Cervantes
Dr. Gabriel Escarela Pérez

Puebla, Pue. Abril 2021

Dedicado a mi familia

Agradecimientos

Gracias a Dios por tener a una gran familia la que siempre me ha apoyado en todas mis decisiones tanto personales como académicas, en especial a mi padre y mi madre José Oscar Méndez Rojas y María Isabel Cuanalo Cordero fundamentales en mi vida ya que sin ellos esto no hubiese sido realidad. Personas muy especiales que siempre confiaron en mí, que no dudaron que lo lograría, que están conmigo hasta el final y que esto también es suyo, el esfuerzo fue de todos y se está logrando. Les doy las gracias por el apoyo infinito, por sus consejos y por su amor.

El camino no fue sencillo, pero con mucho esfuerzo y dedicación se esta logrando, hay que sobreponernos a las adversidades y nunca rendirnos en los sueños que tenemos tanto personales como académicos.

Gracias a mis compañeros que fueron otra parte importante en el transcurso de la carrera estudiantes muy brillantes, pero aún mejor excelentes personas y que algunos se volvieron parte importante en mi vida personal a las cuales aprecio y admiro mucho.

Les agradezco a los Doctores Hugo Adán Cruz Suárez, Bulmaro Juárez Hernández, José Juan Castro Alva, Gabriel Escarela Pérez que gracias a sus consejos sus comentarios y sobre todo su apoyo esto se volvió realidad, excelentes profesores y excelentes personas que siempre me brindaron su apoyo en cualquier momento y me ayudaron a seguir adelante.

Por último, quiero agradecer a la Doctora Hortensia Josefina Reyes Cervantes una profesora excelente siempre me brindo su apoyo, me ayudo a sobre salir, a no renunciar a las adversidades, a seguir luchando por ese sueño, una persona que me tuvo confianza en todos los momentos y que siempre le estaré agradecido por todos sus consejos y su enseñanza tanto personal como académica.

Muchas gracias a todos.

Resumen

El análisis de Curvas ROC es una técnica estadística de decisión, esto es que permite discriminar entre dos grupos o subpoblaciones de una población general partiendo de la medida de una característica en particular, proporcionando un punto de corte a partir del cual se clasifica a los individuos de la población en alguno de los dos grupos de interés.

La curva es un gráfico que resulta de representar, para cada valor umbral, las medidas de sensibilidad y especificidad de la prueba diagnóstica, más adelante se definirán estos conceptos.

En la actualidad el uso de Curvas ROC está muy extendido en diversas ciencias, especialmente en psicología, psiquiatría y medicina, donde se desarrollan diferentes investigaciones científicas en líneas tan relevantes como enfermedad de alzheimer, desorden bipolar, demencia, anorexia nerviosa, evaluación de aprendizaje, memoria, percepción sensorial, detección de señal, desordenes afectivos, ansiedad entre otros.

Este análisis estadístico se aplica generalmente en un sistema compuesto por un dispositivo de recolección de datos y un generador de decisiones (clasificador, diagnosticador o variable de predicción), combinándose con funciones de observación y toma de decisiones basadas en la clasificación o diagnóstico.

Los dispositivos de recolección de datos suelen estar formados por mecanismos de detección o receptores, de donde proviene la denominación Receiver Operating Characteristic.

Por otra parte, el origen de esta metodología tuvo lugar durante la Segunda Guerra Mundial y fue desarrollada por ingenieros electrónicos y estadísticos matemáticos para

determinar si un receptor electrónico es capaz de distinguir satisfactoriamente entre señal y ruido. La estructura conceptual se cimentó en la Teoría de Decisión Estadística y las propiedades estadísticas de las señales y ruidos aleatorios que permitió desarrollar la metodología sobre procesos de detección y reconocimiento de una señal degradada por el ruido.

En particular; las curvas ROC han sido extensamente utilizadas en la Teoría de Detección de Señales para representar la relación entre las fracciones de éxito y de falsas alarmas de clasificadores, lo que permite llevar acabo la toma de decisiones bajo incertidumbre a través de términos precisos y gráficos específicos.

En la actualidad, el análisis de curvas ROC suministra métodos para evaluar un sistema de diagnóstico mediante: un índice fiable y válido de su exactitud o capacidad de discriminación y la valoración de la utilidad de un sistema en términos de costes y beneficios.

Abstract

ROC curve analysis is a statistical decision technique that allows discrimination between two groups or subpopulations of a general population from a characteristic measured on it, providing a cut-off point that classifies individuals in the population into the two interest groups. The curve is the graph resulting from represent, for each threshold value, the measures of sensibility and specificity of the diagnostic prove, later were be define this concepts.

In the actuality the use of the ROC curve is so extended in the different sciences, especially in psychology, psychiatry and medicine, where they stand out scientific researches in so relevant lines like the Alzheimer sick, Bipolar disorder, dementia, anorexia nervosa, learning assessment, memory, sensory perception, signal detection, affective disorders, anxiety among others.

This statistical analysis is generally applicatted in a system composited by a data collection device and a decisions generator (classifier, diagnostician, or prediction variable), combining observation and decision-making functions based on classification or diagnosis. In first place, the data collect devices are usually formed by detection mechanisms or receivers, where the Receiver Operating Characteristic denomination comes from.

On the other hand, the origin of this methodology had place during the second war world and it was developed by engineers and mathematical statisticians for determinate if an electronic receiver was able to distinguish satisfactorily between signal and noise. The conceptual structure was cemented in the statistical decision theory and the statistical properties of the random signals and noises which allowed

develop the methodology about detection and recognition processes of a signal degraded by noise.

In particular the ROC curves have been widely used in the Signal Detection Theory for represent the relation between success fractions and false classifier alarms, allowing decision-making to be carried out under uncertainty through precise terms and specific graphics.

Nowadays the ROC curve analysis provides methods for evaluating a diagnostic system through a reliable and valid index of its accuracy or intrinsic discrimination ability and valuation of the utility of a system in terms of costs and benefits.

Índice general

Índice de figuras

Índice de tablas

Introducción

El área de salud es una parte fundamental en el desarrollo de un país, estado o región y con ello también la toma de decisiones de las distintas autoridades competentes. Es así que en algunos países hospitales, empresas e instituciones privadas y públicas participan en investigaciones enfocadas en esta área, estudiando por ejemplo la propagación de enfermedades infecciosas en áreas y poblaciones que van creciendo en el tiempo, enfermedades que pueden llegar a ser mortales.
La necesidad de interpretar patrones epidemiológicos para implementar programas efectivos de control ha llevado a que científicos y profesionales de la salud, mediante modelos matemáticos, empleén el uso de métodos cuantitativos para estudiar el control de enfermedades infecciosas.

En diversos campos del conocimiento se presenta el problema de la clasificación con base a una o más variables independientes, el área de la salud es un ejemplo donde es una tarea el clasificar a los individuos. Si planteamos la clasificación en dos grupos, a partir de modelos de regresión logística es posible obtener información para cada variable cuyos valores se interpretan como la probabilidad de pertenecer a alguno de ambos grupos. Para estas situaciones, si pensamos en estos dos grupos como sanos y enfermos, existen dos tipos de errores: clasificar a un individuo no enfermo como uno que presente la enfermedad o bien clasificar a una persona enferma como una que no presente la enfermedad. Evitar estos errores corresponden a los conceptos de sensibilidad y especificidad. Por lo que un buen sistema de clasificación sería aquel que maximiza la sensibilidad y especificidad. Aunque

dependiendo de el trabajo del investigador la importancia entre estos tipos de errores puede no ser la misma, con ello cambiaría también la manera de determinar a un buen sistema de clasificación. Un método utilizado para determinar la calidad diagnóstica de una prueba es la conocida curva ROC, siendo así una herramienta para estimar la validez de una prueba diagnóstica, de forma más general permite evaluar la precisión de algunos modelos estadísticos (regresión logística, análisis discriminante) que clasifican a los individuos en 2 categorías.

De manera breve la Curva ROC (Receiver-Operating Characteristic) inició en la Segunda Guerra Mundial como respuesta al problema que planteaban las señales recolectadas por radar es decir el éxito y fracaso para ver si discriminaban de forma correcta. Esto dio inicio al desarrollo de la Teoría de Detección de Señales (Green D.M. (1996)). Poco tiempo después en 1960 empezó a utilizarse en diferentes ramas de estudios experimentales, un ejemplo es en psicofísica por el supuesto de que la energía emitida por seres humanos era reconocible y distinguible de otras señales.
La utilidad más importante y potencial de la Curva ROC en estudios médicos fue propuesta por Lusted (Lusted L.B. (1971)), se deriva de la teoría de detección de señal donde se usa para determinar si un receptor electrónico puede distinguir satisfactoriamente entre señal y ruido, esta metodología ha sido adaptada a diversas áreas clínicas sobresalientes, con un sometimiento de las diferentes pruebas o test diagnósticos, como pueden ser: Laboratorio, Epidemiología, Radiología (diagnóstico por imagen), Bioinformática.
Trazar la curva ROC es una forma popular de mostrar la precisión discriminatoria de una prueba de diagnóstico para detectar si un paciente tiene o no una enfermedad. Las pruebas de diagnóstico médico están diseñadas para discriminar entre diferentes estados de salud, por ejemplo pacientes que están enfermos y los que no lo están (para entender los diferentes conceptos, utilizaremos el término enfermo como una condición que la prueba o test diagnóstico va a detectar, independientemente de que la persona "no enferma" pueda tener otros diferentes problemas de salud). Antes de que las pruebas de diagnóstico se implementen en la práctica, es normal que se estudie su precisión o capacidad para discriminar. Recientemente, el interés ha ido más allá de

determinar la precisión básica de una prueba. Para obtener una mejor comprensión de una prueba, los investigadores están interesados en determinar los factores que afectan su precisión. Al hacerlo, es posible identificar poblaciones y entornos donde una prueba es más o menos precisa, lo que puede ser útil para diferentes usos del autor. Esto se logra mediante el análisis de la precisión de la prueba, que se menciona en este trabajo. La precisión de la prueba de resultados binarios (es decir, positiva o negativa), se resume generalmente con la fracción de verdaderos positivos (FVP) y la fracción de falsos positivos (FFP). La FVP, también llamada sensibilidad, es la proporción de sujetos enfermos detectados correctamente por la prueba. Por otro lado, la FFP o (1-especificidad) se define como la proporción de sujetos no enfermos erróneamente considerados positivos por la prueba. Las curvas características de funcionamiento del receptor (ROC) son una medida de precisión bien aceptada para las pruebas que producen resultados ordinales o continuos. Basado en la noción de usar un umbral o punto de corte para clasificar a los sujetos como positivos o negativos, una curva ROC es un gráfico de FVP versus FFP para todos los puntos de corte posibles. Por lo tanto, describe todo el rango de posibles características para la prueba y, por lo tanto, su capacidad discriminante entre personas enfermas y no enfermas.

Con esto se justifica el dar seguimiento de los últimos desarrollos que se han llevado acabo en un marco teórico, obtener la clasificación con diferentes métodos de estimación propuestos en las últimas publicaciones especializadas a manera de resumen, esto con el objetivo de un mejor manejo conceptual que amplíe las posibilidades de aplicaciones y desarrollo a futuro de una herramienta tan importante en el área de la salud como lo es la curva ROC.

En este trabajo se pretende mostrar las diferentes relaciones entre los métodos de estimación propuestos observando lo que aporta cada uno de ellos, se aplicarán las definiciones con un método de estimación para observar y analizar el posible comportamiento estadístico de un marcador.

Las contribuciones más relevantes de este trabajo son:

Se explica lo esencial de la Curva ROC, métodos de estimación, y se muestra la

relevancia de las Curvas ROC en un caso real. Por ello el objetivo de este trabajo es clasificar a nuestra muestra de sanos y enfermos de manera correcta y se hace mediante la curva ROC, en ella se muestra la sensibilidad y especificidad de nuestro conjunto de datos que es una muestra de 286 observaciones. Para poder calcular el área bajo la curva que indica si estamos haciendo "bien las cosas a la hora de clasificar" y llegar a nuestro objetivo principal, que es encontrar el punto óptimo de nuestra clasificación que se hará mediante el método de Youden (en el cual hablaremos sobre sus ventajas y desventajas), nuestra metodología será utilizando una paquetería de R la cual se llama pROC la utilizaremos para crear la curva ROC y después poder calcular el punto de corte.

Hablaremos sobre que se necesita para tener una curva ROC como lo es una prueba diagnóstica, sensibilidad, especificidad, entre otras cosas se menciona la fiabilidad de las pruebas y de su eficacia y precisión. También hablaremos sobre sus diferentes métodos y su área bajo la curva ROC.

Por último, aplicaremos la teoría en un estudio real a través del software estadístico R. Usando datos reales de una muestra de pacientes a quienes les fueron medidos dos biomarcadores con el objetivo de clasificarlos con presencia o ausencia de cáncer de mama. Formaremos las diferentes curvas ROC y determinaremos aquella que clasifica mejor a los individuos con presencia o ausencia de la condición.

Capítulo 1

Conceptos previos

Definición 1.1. Un espacio de probabilidad es una terna (Ω, F, P), en donde Ω es un conjunto arbitrario, F es una $\sigma - álgebra$ de subconjuntos de Ω, y P es una medida de probabilidad definida sobre F.

Definición 1.2. Una colección F de subconjuntos de Ω es una $\sigma - álgebra$ si cumple las siguientes condiciones:

1. $\Omega \in F$.
2. Si $A \in F$, entonces $A^c \in F$.
3. Si $A_1, A_2, ... \in F$, entonces $\bigcup_{n=1}^{\infty} A_n \in F$.

Definición 1.3. Sea (Ω, F) un espacio medible. Una medida de probabilidad es una función $P : F \to [0, 1]$ que satisface:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0$, para cualquier $A \in F$.
3. Si $A_1, A_2, ... \in F$ son ajenos dos a dos, esto es, $A_n \cap A_m = \emptyset$ para $n \neq m$, entonces $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$.

La probabilidad condicionada es una definición que se une a los axiomas para formar los pilares de la teoría de la probabilidad la cual se denota por $P(A|B)$, mide cómo cambia nuestro conocimiento probabilístico de un suceso A cuando conocemos que otro suceso (digamos, $B \neq \emptyset$) se ha verificado.

El hecho de conocer que el suceso B se ha verificado implica que el experimento aleatorio con el que trabajamos ha cambiado esto se puede expresar como:

$$\epsilon < \Omega, F, P > \xrightarrow{B} \epsilon_B < \Omega_B, F_B, P_B >$$

esto es, partimos del experimento y la verificación de B nos hace pasar a un experimento B en el que las tres entidades involucradas (espacio muestral, clase de sucesos y ley de asignación de probabilidades) han cambiado entonces tenemos

$$P(A|B) = \frac{P(A \cap B)}{P(B)},\ P(B) \neq 0.$$

Esta nueva definición debe ser coherente con los axiomas de la probabilidad. Veamos que, en efecto, así es

1. $P(A|B) = \dfrac{P(A \cap B)}{P(B)} \geq 0$ es el cociente de dos probabilidades y por ellos el cociente de numeros no negativos.
2. $P(\Omega|B) = \dfrac{P(\Omega \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1.$
3. Sea $A_1, A_2, ...$ elementos de F tales que $A_n \cap A_m = \emptyset$ Para todo $n \neq m$.

Veamos que $P\left(\bigcup_{n=1}^{\infty} A_n|B\right) = \sum_{n=1}^{\infty} P\left(A_n|B\right)$.

Dado que $\bigcup_{n=1}^{\infty} A_n \in F$

$$P\left(\bigcup_{n=1}^{\infty} A_n|B\right) = \frac{P\left(\left(\bigcup_{n=1}^{\infty} A_n\right)\bigcap B\right)}{P(B)}$$

$$= \frac{P\left(\bigcup_{n=1}^{\infty}\left(A_n\bigcap B\right)\right)}{P(B)}.$$

Luego, ya que $A_n \cap B \subset A_n$, entonces $(A_n \cap B) \cap (A_m \cap B) = \emptyset$ Para todo $n \neq m$ de modo que $\bigcup_{n=1}^{\infty}\left(A_n\bigcap B\right)$ es la unión infinita de conjuntos ajenos dos a dos, entonces

tenemos

$$P\left(\bigcup_{n=1}^{\infty}\left(A_n\bigcap B\right)\right)=\sum_{n=1}^{\infty}P\left(A_n\bigcap B\right),$$

entonces

$$\begin{aligned}P\left(\bigcup_{n=1}^{\infty}A_n|B\right)&=\frac{\sum_{n=1}^{\infty}P\left(A_n\bigcap B\right)}{P(B)}\\&=\frac{P(A_1\cap B)}{P(B)}+\frac{P(A_2\cap B)}{P(B)}+...\\&=P(A_1|B)+P(A_2|B)+...\\&=\sum_{n=1}^{\infty}P\left(A_n|B\right).\end{aligned}$$

Teorema de probabilidad total 1.1. Sea (Ω, F, P) un espacio de probabilidad, y sea $A_1, A_2, ...$ una partición de Ω tal que cada elemento de la partición es un evento con probabilidad estrictamente positiva.

$$P(B)=\sum_{n=1}^{\infty}P(B|A_n)P(A_n).$$

Teorema de Bayes 1.2. Sea (Ω, F, P) un espacio de probabilidad, y sea $A_1, A_2, ...$ una partición de Ω tal que cada elemento de la partición es un evento con probabilidad estrictamente positiva. Para cualquier evento B tal que $P(B) > 0$, y para cualquier $m \geq 1$ fijo tenemos:

$$P(A_m|B)=\frac{P(B|A_m)P(A_m)}{\sum_{n=1}^{\infty}P(B|A_n)P(A_n)}.$$

Definición 1.4. Una variable aleatoria real es una función $X : \Omega \to \mathbb{R}$ tal que para cualquier Boreliano B se cumple que el conjunto $X^{-1}B$ es un elemento de F.

Proposición 1.1. Una función $X : \Omega \to \mathbb{R}$ es una variable aleatoria si, y sólo si, para cada x en $\mathbb{R}$ se cumple que $\{X \leq x\} \in F$.

Definición 1.5. La función de distribución de una variable aleatoria X es la función $F : \mathbb{R} \to [0, 1]$, definida como sigue

$$F(x) = P(X \leq x),\ x \in \mathbb{R}.$$

Proposición 1.2. Sea $F(x)$ la función de distribución de una variable aleatoria. Entonces

1. $\lim_{x \to +\infty} F(x) = 1$.
2. $\lim_{x \to -\infty} F(x) = 0$.
3. Si $x_1 \leq x_2$, entonces $F(x_1) \leq F(x_2)$.
4. $F(x)$ es continua por la derecha, es decir, $F(x+) = F(x)$

Distribución Bernoulli. Un ensayo Bernoulli es un experimento aleatorio con únicamente dos posibles resultados, llamados genéricamente *éxito* y $fracaso$, y con probabilidades respectivas p y $1 - p$. Se define la variable aleatoria X como aquella función que lleva el resultado éxito al número 1, y el resultado fracaso al número 0. Entonces se dice que X tiene una distribución Bernoulli con parámetro $p \in (0, 1)$. Se escribe $X \sim Ber(p)$ y la correspondiente función de probabilidad es

$$f(x) = \begin{cases} 1 - p, & \text{si } x = 0, \\ p, & \text{si } x = 1, \\ 0, & \text{otro caso.} \end{cases}$$

Definición 1.6. Un parámetro es una caracterización numérica de la distribución de la población de manera que describe, parcial o completamente, la función de densidad de probabilidad de la característica de interés.

Capítulo 2

Clasificación

En general los seres humanos clasificamos todas las cosas que nos rodean de diferentes formas por tamaño, genero, color, edad etc. Es importante mencionar que en estadística hay muchas formas de clasificar y diferentes métodos para realizarlos. Por ejemplo, en temas sociales no hay coincidencia en muchas instituciones en clasificar el inicio y final de la adolescencia. O en temas económicos existe conflicto al asignar una zona con pobreza, en cambio, cuando se estudian objetos mediante un proceso industrial o mecánico es más fácil hacer clasificaciones tomando en cuenta las medidas o controles de calidad asignadas por la empresa. Ahora nos centraremos en la clasificación en el área de medicina en donde los científicos clasifican a las personas mediante distintas pruebas para un mejor resultado, de esta forma obtenemos mucha más información acerca de los individuos a través de dichas pruebas.

En este capítulo hablaremos de la precisión de las pruebas.

2.1 Test diagnóstico

Las pruebas a pacientes se utilizan para clasificar, éstas son llamadas test diagnósticos o pruebas diagnósticas. Se definen como procedimientos aplicados con la finalidad de detectar una "condición" medica determinada en el individuo, el término "condición"

se puede referir a una enfermedad, un síndrome, o un proceso patológico.

La complejidad para distinguir la "condición" generalmente no es susceptible de ser observada directamente, por lo cual se hacen estudios diagnósticos que proveén información para así clasificar al sujeto. Este clasificador comúnmente llamado marcador o biomarcador proporciona un "resultado" y puede ser:

• Binario: Cuando solo existen dos posibles resultados, positivo (P) y negativo (N) (Por ejemplo: cuando una mujer se hace una prueba de embarazo, el resultado puede ser positivo o negativo).
• Ordinal: Cuando el resultado puede tomar distintos valores ordenados siguiendo una escala establecida (Por ejemplo: en diferentes tipos de pruebas (académicas) la mayoría de veces los resultados pueden ser de la siguiente forma: malo, regular, bueno, excelente).
• Continuo: Cuando el resultado puede ser cualquier valor dentro de un intervalo real especificado (Por ejemplo: se va a realizar un estudio químico y se revisan los valores del colesterol).
Se va a denotar a la condición real del sujeto o estatus de la enfermedad por D (del inglés Disease).
En general un clasificador tiene dos propósitos:
• Proporcionar información fiable sobre el estado o condición de un individuo.
• Influir en la acción apropiada para el estado pronosticado de un individuo.

2.2 Validez de las pruebas diagnósticas

Una prueba diagnóstica será válida si es capaz de medir correctamente el fenómeno que pretende estudiar, para que nosotros podamos evaluar la validez de una prueba diagnóstica se necesita un patrón de referencia que refleje la característica a medir, para el caso más simple, consiste en medir la presencia o ausencia de una enfermedad, por lo cual el patrón de referencia tendrá que clasificar perfectamente a la población

enferma y a la sana. Generalmente aunque asumamos que el patrón de referencia tiene una validez absoluta, es frecuente que no sea perfecta.

En ocasiones no tenemos ninguna referencia de la prueba, por la complejidad del concepto a medir o por la ausencia de conocimiento. En estas situaciones resulta útil recurrir a criterios diagnósticos que estan diseñados por científicos o resultados de un conjunto de pruebas agrupadas.

2.3 Resultados de un test diagnóstico

El verdadero estado de un individuo está determinado por un gold standard (patrón de oro). Ahora sea la variable aleatoria D=estado que sigue una distribución Bernoulli de parámetro p que llamaremos prevalencia del evento sobre la población, de modo que dicha variable es como sigue:

$$D\sim\text{Ber(p)}\Rightarrow D=\begin{cases}0, & \text{si } \textit{no presenta el evento de interés el individuo}\\ 1, & \text{si } \textit{presenta el evento de interés el individuo.}\end{cases}$$

Lo cual representa el resultado de un test diagnóstico o marcador binario.
En este contexto, la variable dicotómica determinada gold standard establece la presencia de una determinada condición, es decir, representar el verdadero estado del individuo.

Definición 2.1. Dada una población, la prevalencia es la proporción de personas que presentan una característica determinada.
A continuación presentaremos la prevalencia en términos probabilísticos

$$\textit{prevalencia} = p = P(\textit{estado} = 1) = P(D = 1) = P(\textit{presentar el evento}),$$

por otra parte tenemos el complemento

$$1 - p = P(\textit{estado} = 0) = P(D = 0) = P(\textit{no presentar el evento}).$$

Los elementos que necesitamos son:
• Una muestra aleatoria simple que no presente el evento de interés, a los que llamaremos sanos y una muestra aleatoria simple que si lo presente a los que llamaremos enfermos.

• Se requiere una variable aleatoria X que mida cierta característica en cada individuo, cuyo resultado puede ser continuo o discreto.
•Una variable aleatoria Bernoulli que llamaremos $Y = \textit{prueba}$ es la que queremos estudiar su eficiencia discriminante que tomará dos resultados positivo o negativo en

función del valor de X respecto al que denominaremos punto de corte o valor umbral c. Dicho de otra forma las respuestas del clasificador indicando la presencia o aunsencia de la condición se denotan por:

$$Prueba = \begin{cases} Positivo = Y = 1, & \text{si } c \leq X \\ Negativo = Y = 0, & \text{si } c > X. \end{cases}$$

Observación. El verdadero estado de la condición no tiene por qué coincidir con el resultado de la clasificación proporcionada por el clasificador, excepto en el caso de un clasificador ideal o perfecto.
Si extraemos una muestra directamente de la población, un estimador de la prevalencia es la razón:

$$p = \frac{número\ de\ enfermos\ de\ la\ muestra}{cantidad\ de\ individuos\ de\ la\ muestra} \quad .$$

Si la muestra es heterogénea este es un mal estimador, por lo que la literatura sugiere extraer tanto enfermos en función de sanos, como sanos en función de enfermos.

2.3.1 Matriz de confusión

La matriz de confusión es una matriz de orden 2, donde las filas representan las clases estimadas y las columnas las clases actuales. Con esta forma aunque es simple es muy efectiva ya que se pueden visualizar los aciertos y errores que ha cometido el clasificador. La suma de valores en la diagonal principal representa las decisiones acertadas del clasificador, y la diagonal secundaria representa las decisiones equivocadas del clasificador.
A continuación se presenta en la Tabla 2.1.

	$Enfermo = D = 1$	$Sano = D = 0$
$Prueba+ = Y = 1$	$Verdadero\ positivo(V_+)$	$Falso\ positivo(F_+)$
$Prueba- = Y = 0$	$Falso\ negativo(F_-)$	$Verdadero\ negativo(V_-)$

Tabla 2.1: Matriz de confusión.

Escribiendo de otra forma lo anterior, tenemos como resultado la Tabla 2.2

$V_+\ si\ (D = 1, Y = 1)$
$F_+\ si\ (D = 0, Y = 1)$
$F_-\ si\ (D = 1, Y = 0)$
$V_-\ si\ (D = 0, Y = 0)$

Tabla 2.2: Otra forma de expresar la matriz de confusión.

Los valores de la diagonal principal denominada diagonal de aciertos tiene como elementos:

• Verdaderos positivos (V_+): el número de casos con presencia de la condición y predicción correcta, es decir, el número de casos con presencia de la condición que son clasificados como positivos correctamente.

• Verdaderos negativos (V_-): el número de casos con ausencia de la condición y predicción correcta, es decir, el número de casos con ausencia de la condición que son clasificados como negativos correctamente.

Asimismo los valores que aparecen fuera de la diagonal de aciertos representan los resultados incorrectos de la clasificación (errores o confusión) los elementos son:

• Falsos positivos (F_+): el número de casos en los que está ausente la condición y con predicción incorrecta, es decir, el número de casos con ausencia de la condición que son clasificados como positivos erróneamente.

• Falsos negativos (F_-): el número de casos en los que está presente la condición y con predicción incorrecta, es decir, el número de casos con presencia de la condición que son clasificados como negativos erróneamente.

Estos resultados erróneos pueden provocar graves consecuencias y riesgos para el individuo. Así, una falsa detección (el clasificador indica falsamente la presencia de la condición en un individuo cuyo verdadero estado es ausencia de la condición) conlleva acciones innecesarias e incorrectas. Por otro lado, los errores del tipo falsos negativos derivan en impedir o atrasar las acciones necesarias y correctas a un individuo que realmente presenta la condición, pero sin embargo el clasificador indica falsamente la ausencia de la misma.

Además, las sumas por columnas o por filas de la matriz de confusión corresponden a:

• $TCP = V_+ + F_-$: Total de respuestas con presencia de la condición de interés.
• $TCA = F_+ + V_-$: Total de respuestas con ausencia de la condición de interés.
• $TRP = V_+ + F_+$: Total de respuestas positivas.
• $TRN = F_- + V_-$: Total de respuestas negativas.

En la Figura 2.1 se presentan unas distribuciones de sanos y enfermos, cabe esperar que dichas distribuciones sean ajenas entre sí, cuanto menos solapamiento mejor resultado obtendremos, es decir, los falsos negativos y positivos tendrán un valor mucho menor y teóricamente es lo que necesitamos. Esta separación no dependerá de la capacidad discriminante de la prueba.

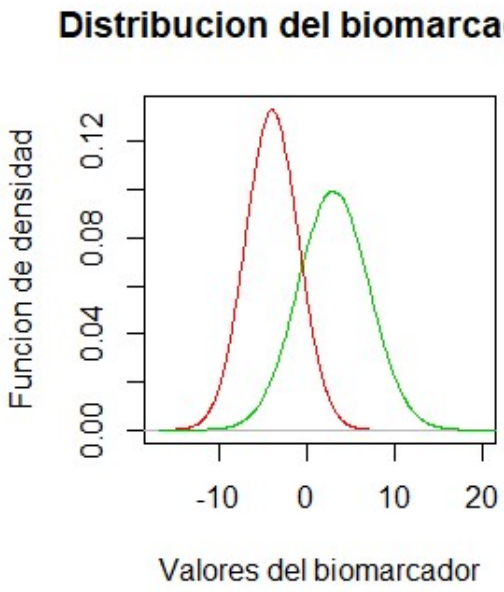


Figura 2.1: Distribución de enfermos (curva verde) y sanos (curva roja).

(Torres A. (2010)) dice que Yerushalmy en 1947 introduce los términos de sensibilidad y especificidad como indicadores estadísticos que evalúan el grado de eficacia inherente a una prueba diagnóstica.

2.3.2 Sensibilidad

La sensibilidad es un parámetro que se mide en el grupo de sujetos que verdaderamente están enfermos, se puede observar como la probabilidad de clasificar correctamente a un individuo cuyo estado es la presencia de la condición de interés.
Dicho de otra forma es la probabilidad que si tenemos un individuo enfermo, que la prueba lo clasifique como enfermo.

$$S = P(Y = 1|D = 1) = \frac{P(\{Y = 1\} \cap \{D = 1\})}{P(D = 1)} \quad .$$

La definicion es poblacional y se puede representar con un estimador muestral.
Su estimador es la proporción de respuestas positivas que estan bien clasificadas, queda de la siguiente manera:

$$S = \frac{Enfermos\ positivos}{Total\ enfermos} = \frac{V_+}{V_+ + F_-} \quad .$$

O bien, fracción de verdaderos positivos (FVP). El término proporción puede causar confusión en epidemiología resultando más apropiado sustituirlo por fracción.
Si la sensibilidad del clasificador es máxima, S=1 entonces $F_- = 0$ equivalentemente $V_+ = TCP$ (total de respuestas con presencia de la condición) ya que $TCP = V_+ + F_-$.

En situaciones de diagnóstico médico, si las consecuencias son graves en caso de no detectarse a tiempo, resulta fundamental llevar a cabo procedimientos con el mínimo F_- posible. Por ello, para utilizar nuevas técnicas de detección se debe elegir aquella con máxima sensibilidad.
Así mismo, aplicando el complemento obtenemos

$$1 - S = \frac{V_+ + F_-}{V_+ + F_-} - \frac{V_+}{V_+ + F_-} = \frac{F_-}{V_+ + F_-},$$

conocida como la fracción de falsos negativos (FFN).

2.3.3 Especificidad

La especificidad es un parámetro que se mide en el grupo de sujetos no enfermos, se puede observar como la probabilidad de clasificar correctamente a un individuo cuyo estado es la ausencia de la condición de interés.
Dicho de otra forma es la probabilidad de que si un individuo sano, que la prueba lo clasifique como sano.

$$E = P(Y = 0|D = 0) = \frac{P(\{Y = 0\} \cap \{D = 0\})}{P(D = 0)} \quad .$$

Siguiendo el razonamiento utilizado en la sensibilidad obtenemos:

$$E = \frac{Sanos\ negativos}{Total\ sanos} = \frac{V_-}{V_- + F_+},$$

conocida como la fracción de verdaderos negativos(FVN).

Si la especificidad es máxima, E=1, entonces $F_+ = 0$, equivalentemente $V_- = TCA$ (total de respuestas con ausencia de la condición), puesto que $TCA = F_+ + V_-$.

En situaciones como en el caso de una enfermedad terminal en una etapa avanzada, puede resultar muy grave comunicar al paciente que la padece cuando en realidad esta sano, por lo que se requiere máxima especificidad para minimizar los falsos positivos.
Ahora, aplicando el complemento obtenemos

$$1 - E = \frac{V_- + F_+}{V_- + F_+} - \frac{V_-}{V_- + F_+} = \frac{F_+}{V_- + F_+},$$

llamada fracción de falsos positivos (FFP).

Así pues, un clasificador perfecto sería aquel que no comete errores de clasificación, esto es, $F_- = F_+ = 0$, matriz de confusión diagonal.
Para individuos con resultados positivos en la prueba, dicha probabilidad a posteriori (es la probabilidad condicional que es asignada después de que la evidencia es tomada en cuenta.) (Passas M. (2012)) es:

$$P(D = 1|Y = 1) = \frac{P(Y = 1|D = 1)P(D = 1)}{P(Y = 1|D = 1)P(D = 1) + P(Y = 1|D = 0)P(D = 0)}$$

$$= \frac{Sp}{Sp + (1 - E)(1 - p)} \ .$$

Cuya estimación muestral se calcula mediante el valor predictivo positivo o precisión de un clasificador, es la probabilidad de identificar correctamente los casos positivos, siendo su estimador

$$VPP = \frac{V_+}{V_+ + F_+} \ .$$

Y, para individuos con resultados negativos en la prueba se tiene:

$$P(D = 0|Y = 0) = \frac{P(Y = 0|D = 0)P(D = 0)}{P(Y = 0|D = 0)P(D = 0) + P(Y = 0|D = 1)P(D = 1)}$$

$$= \frac{E(1 - p)}{E(1 - p) + (1 - S)\ p}.$$

El valor predictivo negativo de un clasificador, es la probabilidad de identificar correctamente los casos negativos, siendo su estimador

$$VPN = \frac{V_-}{V_- + F_-} \ .$$

Para ilustrar estos conceptos utilizamos un ejemplo sobre el diagnóstico de alcoholismo (Bean P. (2000)).

Bean P. (2000), motiva la predicción de pacientes alcohólicos debido a su baja prevalencia en Estados Unidos, del 5 al 7% lo que le asocia a pruebas diagnósticas con capacidad de predicción aceptable.

En este caso, se considera un test para diagnosticar dicho síntoma con una sensibilidad del 65% y una especificidad del 95%. La Tabla 2.3 presenta la clasificación correspondiente a esta prueba diagnóstica, considerando que la prevalencia es igual a 7%.

	Respuesta del test	
	Positivo	Negativo
Es alcohólico	46	24
No es alcohólico	47	883
Total	93	907

Tabla 2.3: Matriz de confusión sobre alcoholismo.

Con los datos de esta tabla podemos obtener los siguientes valores predictivos positivo y negativo, repectivamente:

$$VPP = \frac{46}{93} = 0.49 \text{ y } VPN = \frac{883}{907} = 0.97$$

El valor predictivo positivo del 50% (aproximadamente) indica que existe un 50% de probabilidad de que la respuesta positiva del test corresponda verdaderamente a un individuo alcohólico, y consecuentemente, un 50% de probabilidad de que la respuesta positiva del test corresponda incorrectamente a un individuo no alcohólico.

Este valor predictivo positivo del test, debido a la baja prevalencia de la condición, causa que esta prueba diagnóstica se considere inadecuada para clasificar correctamente a los individuos de la muestra en alcohólicos y no alcohólicos.

Nota: Cuando más cercanas a 1 sean estas probabilidades tendremos mejor capacidad discriminante de nuestra variable de decisión.

Con todo lo anterior podemos construir la siguiente Tabla 2.4

	$Y = 1$	$Y = 0$
$D = 1$	$FVP = \frac{V_+}{V_+ + F_-}$	$FFN = \frac{F_-}{V_+ + F_-}$
$D = 0$	$FFP = \frac{F_+}{V_- + F_+}$	$FVN = \frac{V_-}{V_- + F_+}$

Tabla 2.4: Fracciones de todos los posibles casos.

En otras palabras lo que obtenemos de dicha Tabla 2.4 es:
Probabilidad de obtener un resultado negativo cuando el sujeto no tiene la enfermedad $\frac{V_-}{V_- + F_+}$.

Probabilidad de obtener un resultado positivo cuando el sujeto tiene la enfermedad $\frac{V_+}{V_+ + F_-}$.
Proporción de resultados válidos entre los resultados negativos de la prueba, es decir, es la probabilidad de identificar correctamente los casos negativos $\frac{V_-}{V_- + F_-}$.
Proporción de resultados válidos entre los resultados positivos de la prueba, es decir, es la probabilidad de identificar correctamente los casos positivos $\frac{V_+}{V_+ + F_+}$.
Proporción de individuos sin la condición que son incorrectamente clasificados como positivos, dada por: $\frac{F_+}{V_- + F_+}$.
Proporción de individuos con la condición que son incorrectamente clasificados como negativos, dada por: $\frac{F_-}{V_+ + F_-}$.

Para ilustrar estas medidas de la exactitud de un clasificador consideremos un ejemplo basado en la presencia de cáncer de pecho en una muestra de 60 pacientes dado en (Zhou X. (2002)). La muestra consta de 30 pacientes con cáncer de pecho probado y 30 pacientes con mamografías normales, cuyos resultados estan en la Tabla 2.5.

La pregunta natural que se hace es ¿cuál es la capacidad de una mamografía para discriminar correctamente entre la presencia y ausencia de cáncer?.

	Resultado de la mamografía	
	Positivo	Negativo
Presencia cáncer	29	1
Ausencia cáncer	19	11
Total	48	12

Tabla 2.5: Matriz de confusión sobre 60 mamografías.

De los 30 pacientes con cáncer de pecho, para 29 la mamografía ha resultado positiva, es decir, el radiólogo les indicaba realizar otra prueba confirmatoria. Por tanto, resultan 29 verdaderos positivos y 1 falso positivo, siendo la sensibilidad de la mamografía:

$$S = \frac{29}{29+1} = 0.9667.$$

Por otro lado, de los 30 pacientes sin cáncer de pecho, la mamografía resulta negativa para 11, siendo la especificidad:

$$E = \frac{11}{19+11} = 0.3667, \text{ equivalentemente, } FFP = 1 - E = 0.6333.$$

Por tanto la mamografía detecta correctamente el cáncer de pecho el 96.67% de las veces, aunque el 63.33% de las veces provoca erróneamente una segunda prueba.

Dadas las graves consecuencias de la presencia de cáncer de pecho, es deseable que tenga una mínima fracción de falsos negativos en este caso seria:

$$FFN = 1 - S = 0.0333.$$

2.4 Relación entre sensibilidad y especificidad

Como ya hemos dicho en un test las personas sanas y enfermas siguen distribuciones diferentes y puede que las distribuciones tengan puntos en común.
Ahora cometer un error (Falso Negativo), es decir, fallar al pronosticar la enfermedad cuando esta presente, tiene como resultado personas enfermas sin el tratamiento con posibilidades de morir.
Ahora cometer un error (Falso Positivo), es decir, fallar cuando la persona no tiene la enfermedad conlleva a aplicarle tratamientos a personas sanas (será menos riesgo de muerte).
Con esto se plantea un gráfico con los ejes coordenados 1- Especificidad y Sensibilidad, donde la pareja (1-Especificidad, Sensibilidad) puede tener una precisión perfecta si toma el valor de (0,1). El gráfico obtenido se denomina Curva ROC.

2.5 Errores en el estudio de test diagnósticos

En el ámbito clínico siempre se quiere conocer que tan válidas son las pruebas diagnosticadas asumiendo las características de la prueba. Sin embargo muchas veces los estudios están mal realizados o mal diseñados y por lo tanto obtenemos conclusiones erróneas sobre el trabajo de interés.

Cuando analizamos una prueba diagnóstica nos debemos cuestionar lo siguiente:
¿Ha sido comparada la prueba con un verdadero patrón de referencia (gold standard)? (Ochoa C. (1960)).

Gold standard o patrón de referencia es un término utilizado para definir aquellas pruebas diagnósticas que tienen fiabilidad de diagnosticar una determinada enfermedad. Es importante considerar si el patrón de referencia es capaz de clasificar el estado de enfermedad en todas las observaciones, pero puede ocurrir que haya observaciones con un diagnóstico indeterminado, si son excluidas del análisis se producirán estimaciones sesgadas de las características de las pruebas diagnosticadas, este error se conoce como sesgo por exclusión de indeterminados y ocasiona sobrestimaciones en la sensibilidad y en la especificidad. Otro sesgo, relacionado con el patrón de referencia, que debe tratar de evitarse y es conocido como el sesgo de incorporación. Este sesgo ocurre cuando elementos de la prueba diagnóstica forman parte del patrón de referencia (Ochoa C. (1960)).

Para valorar esta cuestión es preciso que los criterios de selección y las características clínicas y epidemiológicas de la muestra analizada estén claramente presentados. En

el diseño del estudio se debe tratar de garantizar que no se haya eliminado o excluido una observación (paciente) en función del resultado de la prueba o de la existencia de mayor o menor riesgo de enfermedad.

¿Se ha evitado el sesgo de secuencia o verificación diagnóstica? (Ochoa C. (1960)).

El diseño del estudio debe tratar de garantizar que a todos los sujetos se les haya realizado, tanto la prueba diagnóstica, como el patrón de referencia, se van a calcular directamente de los datos siempre que en la muestra no se hayan excluido pacientes, en función del resultado de la prueba o de la existencia de mayor o menor riesgo de enfermedad. Pero esta estrategia simultánea, sin duda la más válida, resulta en ocasiones poco factible.

Una opción más eficiente para la evaluación de pruebas diagnósticas es el diseño retrospectivo. En él, se determina en un primer paso la presencia o ausencia de enfermedad y en un segundo paso se realiza la prueba diagnóstica a dos submuestras representativas de los sujetos con y sin enfermedad. Con esta estrategia podemos calcular directamente la sensibilidad y la especificidad, pero los valores predictivos deben ser obtenidos con las fórmulas bayesianas, que trabajan con probabilidades condicionales (Ochoa C. (1960)).

¿Cuáles son los resultados? (Ochoa C. (1960)).

El objetivo de la realización de la prueba diagnóstica es, una vez conocido el resultado, modificar la probabilidad preprueba (la probabilidad de que un sujeto posea una condición o enfermedad antes de realizar una o varias pruebas.) hasta obtener una probabilidad postprueba (la probabilidad de que un sujeto posea una enfermedad después de realizar la prueba).

Generalmente los resultados se expresan a partir de la proporción de aciertos de la prueba diagnóstica entre las poblaciones enfermas (sensibilidad) y sanas (especificidad).

Otro aspecto importante en la presentación de los resultados es la correcta utilización de las distintas herramientas disponibles para el análisis de la validez de las pruebas diagnósticas. En este sentido, interesa destacar la gran utilidad práctica que tienen herramientas como los cocientes de probabilidades y las curvas ROC.

2.5.1 Fiabilidad de las pruebas diagnósticas

Solo hemos hablado de la validez de las pruebas diagnósticas, pero la calidad de una prueba no depende solamente de su validez sino también de su fiabilidad.
La fiabilidad de una prueba es su capacidad para producir los mismos resultados cada vez que se aplica en condiciones similares. La fiabilidad implica falta de variabilidad, sin embargo las mediciones realizadas por las pruebas diagnósticos están sujetas a múltiples fuentes de variabilidad. A la hora de analizar y controlar la fiabilidad de las

pruebas diagnósticas tienen especial interés en estudiar la variabilidad encontrada entre las mediciones realizadas por dos o más observadores o instrumentos, y la variabilidad encontrada entre mediciones repetidas realizadas por el mismo observador o instrumento.

Capítulo 3

Curva ROC

La curva ROC es una representación gráfica del rendimiento de un clasificador, proporcionando una herramienta visual para examinar la relación entre la capacidad del clasificador para detectar correctamente los individuos con presencia de la condición de interés y los de ausencia.
Zhou (2002) dice las ventajas de un gráfico ROC:

• Es independiente de la prevalencia, dado que está totalmente determinada por los pares de sensibilidad y especificidad de un clasificador.

• Proporciona una comparación visual de dos o más clasificadores, permitiendo dicha evaluación en todos los criterios de clasificación posibles.

Además otra ventaja de la curva ROC es que no requiere seleccionar un punto de corte, pues consiste en una representación que los incluye a todos, no obstante las curvas ROC permiten desarrollar mecanismos de búsqueda y selección de puntos de corte.

El espacio ROC consiste en una serie de coordenadas donde se representa la fracción de falsos positivos (complementario de la especificidad) en el eje de abscisas frente a la fracción de verdaderos positivos (sensibilidad) en el eje de las ordenadas.

Así, el espacio ROC permite visualizar el rendimiento de un clasificador mediante representaciones bidimensionales a partir de los puntos (1-especificidad, sensibilidad). En general el espacio ROC muestra las correspondencias relativas entre beneficios (verdaderos positivos) y costes (falsos positivos) de una clasificación.

En particular, algunos de los puntos del espacio ROC se consideran de gran interés, ya que representan rendimientos extremos. Por ejemplo en la Figura 3.1 representa un espacio ROC, el punto (0,0) nos dice que no representa repuestas positivas, por lo que no se cometen errores del tipo falsos positivos, sin embargo, tampoco se detectan los verdaderos positivos.

El vértice opuesto (1,1) representa la tendencia contraria, es decir, proporciona incondicionalmente respuestas positivas, pero en este caso no se detectan los verdaderos negativos.

El punto (0,1) representa la clasificación perfecta dado que no se cometen errores del tipo falso postivo ni del tipo falso negativo.

Un elemento importante en el espacio ROC es la diagonal que une los puntos (0,0)

y (1,1), utilizada como linea de referencia, la cual representa la tendencia a clasificar aleatoriamente un estado de la condición.

Un clasificador que supone la presencia de la condición en el 40% de los individuos, se puede esperar que clasifique correctamente el 40% de los que presentan la condición, pero tamnbién tendrá un 40% de falsos positivos, esto es, clasificará correctamente el 60% de los que no presentan la condición.

Cualquier punto en la región triangular inferior tendrá peor rendimiento que un clasificador aleatorio. No obstante el espacio es simétrico, respecto a la diagonal que separa ambas regiones triangulares.

Si invertimos el sentido de la predicción, es decir, las respuestas positivas se consideran negativas y viceversa, entonces los errores falsos positivos (negativos) se transforman en verdaderos positivos (negativos). Por tanto puede invertirse la positividad de cualquier punto del espacio ROC situado en la región triangular inferior para producir un punto en la región triangular superior.

De este modo, los puntos que aparecen a la izquierda en la región triangular superior del espacio ROC se consideran estrictos, cometen pocos errores de tipo falso positivo pero también suelen presentar una baja fracción de verdaderos positivos, ya que presentan una débil tendencia a proporcionar respuestas positivas.

Los puntos que se sitúan a la derecha de la región triangular superior del espacio ROC, se cosideran tolereantes, clasifican casi a todos los positivos correctamente, pero también suelen tener una alta fracción de falsos positivos, ya que muestran una fuerte tendencia a proporcionar respuestas positivas.

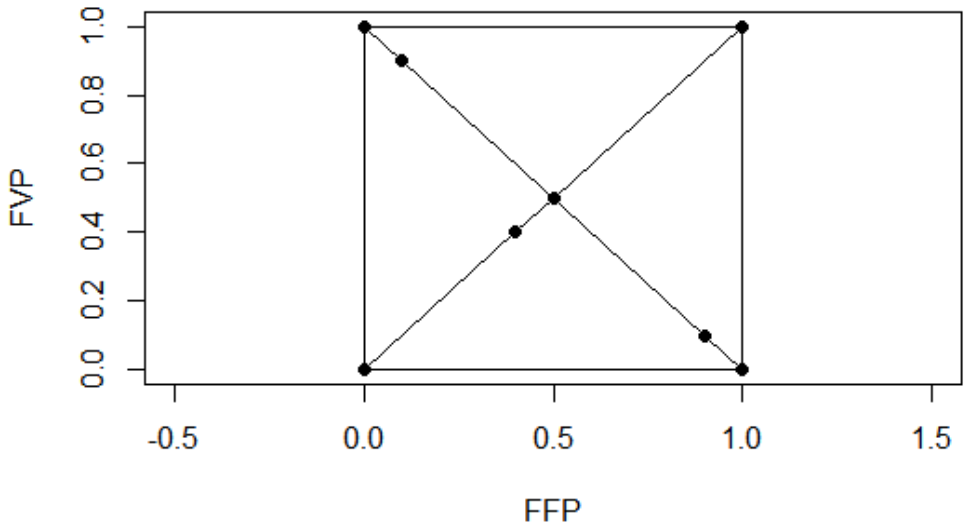


Figura 3.1: Representación del espacio ROC.

Definición 3.1. La curva ROC poblacional representa 1- especificidad frente a la sensibilidad para cada valor umbral o punto de corte en la escala de resultados de la prueba de estudio. Es decir $y = f(x)$ donde

$$ROC(c) = \begin{cases} y = S(c) \\ x = 1 - E(c). \end{cases}$$

Sin embargo ante la dificultad de obtener datos poblacionales podemos aproximarla por la curva ROC muestral, que representa la fracción de falsos positivos y fracción de verdaderos postivos

$$ROC(c) = \begin{cases} y = FVP(c) \\ x = FFP(c). \end{cases}$$

Proposición 3.1. Sean las variables $X_E = (X|D = 1)$ y $X_S = (X|D = 0)$, la variable aleatoria de decisión condicionada al grupo de enfermos y por otro lado, representan al grupo de sanos. Sus correspondientes funciones de distribución son: $F_E(x) = P(X_E \leq x)$ y $F_S(x) = P(X_S \leq x)$ respectivamente. Fijando un evento de interés y tomando a aquellos que son mayores que un valor x, se define la curva ROC asociada a la variable x como la función

$$ROC(t) = 1 - F_E(F_S^{-1}(1 - t)) \qquad 0 \leq t \leq 1$$

donde t es el complemento de la especificidad.

Demostración.
A partir de la definición de la especificidad

$$E = P(y = 0|D = 0)$$

entonces

$$F_S(u) = P(X_S \leq u) = P(X \leq u|D = 0) = E(u).$$

De donde

$$(1 - E)(u) = 1 - F_S(u).$$

Por otro lado a partir de la definición de sensibilidad

$$S = P(y = 1|D = 1)$$

tenemos

$$S(u) = P(X > u|D = 1) = P(X_E > u) = 1 - F_E(u).$$

Ahora, para cada t la curva ROC representa el par de parejas de la forma:

$$(1-E, S) = (1-F_S(u), 1-F_E(u))$$

$$\Rightarrow t = 1-F_S(u)$$

$$\Rightarrow u = F_S^{-1}(1-t)$$

$$\Rightarrow ROC(t) = 1-F_E(F_S^{-1}(1-t)).$$

Para ilustrar esta definición se hace un ejemplo donde X_p y X_n siguen distribuciones normales en ambos grupos con tres puntos de corte diferentes (X_p es muestra de positivos y X_n es muestra de negativos). Se considera X_n=N(-1,3) y X_p=N(3,4).

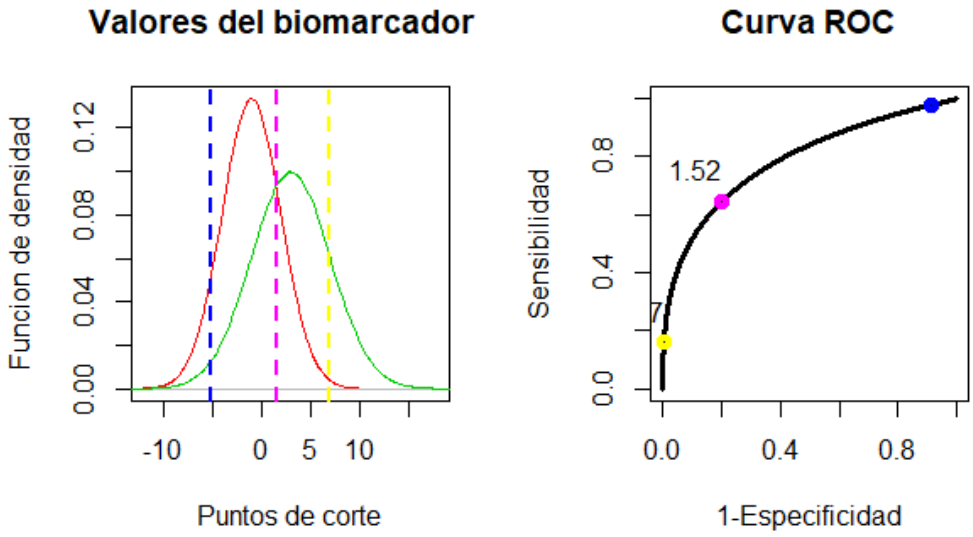


Figura 3.2: Ejemplo curva ROC.

En la Figura 3.2 se muestran tanto las distribuciones de sanos (curva roja) como la de enfermos (curva verde). Se puede observar que ambas distribuciones tienen puntos en común, por lo que tomemos cualquier punto de corte no todos los individuos estarán bien clasificados. Se han tomado tres puntos de corte (-5, 1.5,7), también se observa la curva ROC para los tres puntos de corte.
Podemos observar que tomando el punto de corte -5 la sensibilidad es muy grande (el área es muy pequeña para la distribución de enfermos (curva verde) y esto es la

proporción de casos mal clasificados 1- sensibilidad) mientras que si tomamos el punto de corte 7 la sensibilidad es muy pequeña.
Ahora se observa que la especificidad del punto de corte -5 es muy pequeña debido a que tomar el punto de corte 7 el área es muy pequeña para la distribución de sanos (curva roja) de igual forma están mal clasificados (1-especificidad).
Ahora tenemos que el punto de corte que separa ambas distribuciones 1.5, parece que equilibra la sensibilidad y especificidad.
Por lo que se concluye:
• Tomando como punto de corte -5 la prueba es muy sensible por lo que daría la importancia al clasificar correctamente a los enfermos, permitiendo que los sujetos que no tengan la enfermedad los clasificara como enfermos.
• Tomando el punto de corte 7 la prueba es muy específica por lo que daría importancia al clasificar correctamente a los sanos, permitiendo que los sujetos que tiene la enfermedad los clasificara como sanos.
• Tomando el punto de corte 1.5 la prueba es equilibrada dando la misma importancia al hecho de clasificar correctamente a los individuos sanos como a los enfermos.

La curva ROC permite describir que tan separadas están las distribuciones por ejemplo si las funciones de densidad se cruzan la curva ROC resultante será similar a la mostrada en la Figura 3.4, si las distribuciones se solapan casi por completo, entonces la curva ROC no aporta información acerca de lo que se esté estudiando (algunos autores difieren acerca del aporte que da la curva ROC) y está representada por la línea diagonal como se muestra en la Figura 3.5. Ahora si las funciones están simétricamente distante como en la Figura 3.3 entonces la curva ROC será la línea que este cerca del punto (0,1) indicando un test perfecto, en otras palabras, esta curva ROC es la que mayor información aporta aunque algunos autores no lo apoyan (Fawcett T. (2006)).

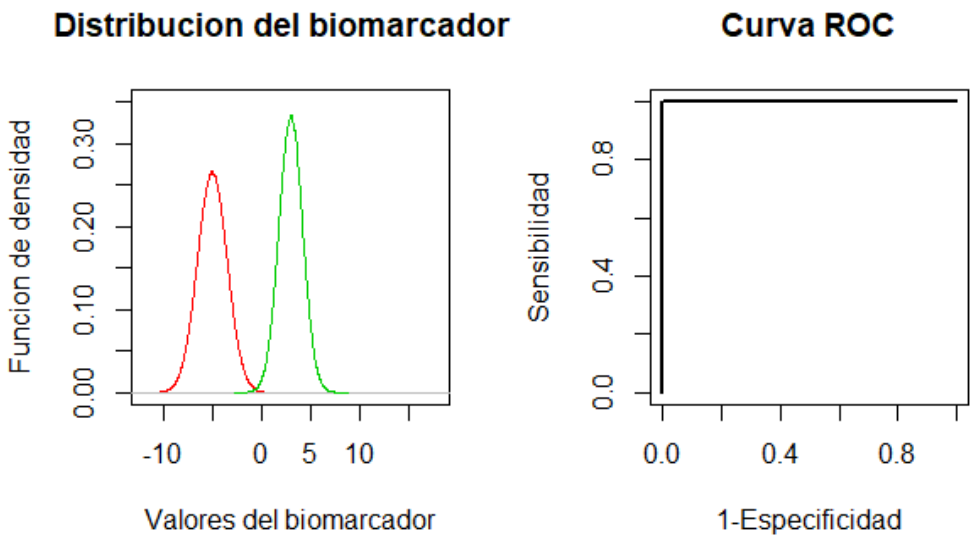


Figura 3.3: Curva ROC de dos poblaciones que no están solapadas.

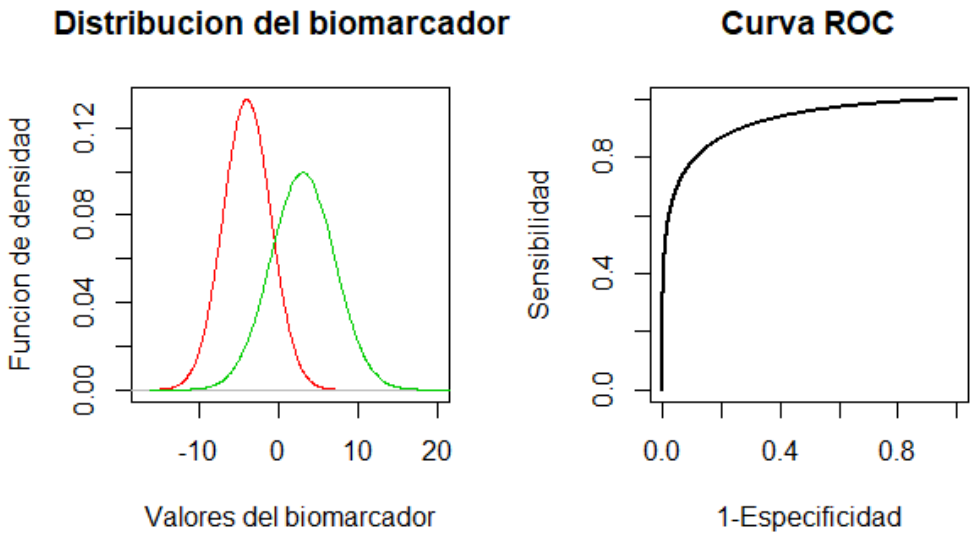


Figura 3.4: Curva ROC de dos poblaciones que no están completamente solapadas.

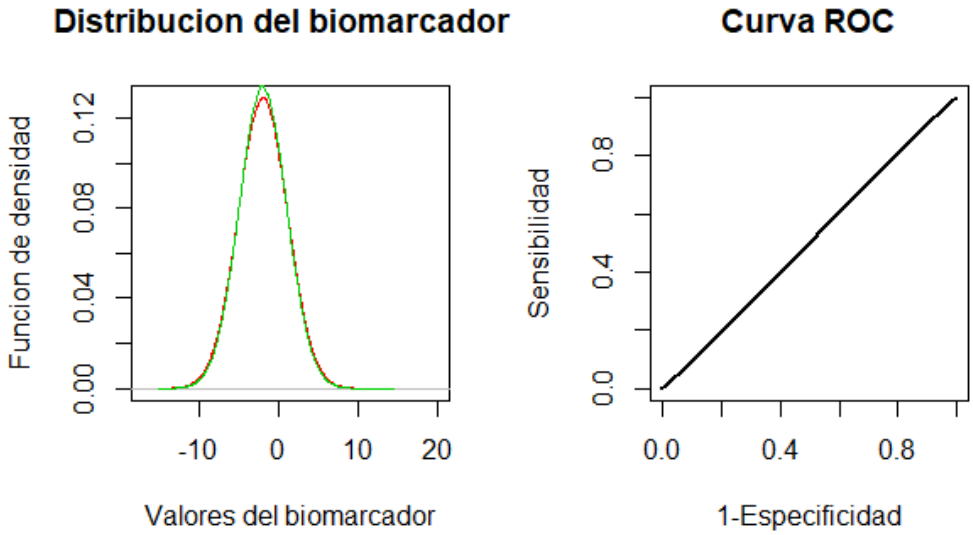


Figura 3.5: Curva Roc de dos poblaciones totalmente solapadas.

Como se ha mencionado la sensibilidad y especificidad son estimaciones, por supuesto queremos ver que tan cerca estamos del valor real, por lo que para un tratamiento estadístico se deben construir intervalos de confianza (se encontrará un rango de valores del parámetro, con una probabilidad determinada). Aquí se usará el método de proporciones (método clasico) puesto que la sensibilidad y la especificidad lo son. Por lo tanto, tenemos como estadístico:

$$Z = \frac{\hat{p} - p}{SD(p)}$$

siendo p la proporción a estimar y $SD(p)$ su desviación estándar.
Ahora con la sensibilidad tenemos:

$$SD(S) = \sqrt{\frac{\frac{V_+}{V_+ + F_-}\ \frac{F_-}{V_+ + F_-}}{V_+ + F_-}} = \sqrt{\frac{V_+\ F_-}{(V_+ + F_-)^3}}\ .$$

Ahora con la especificidad tenemos:

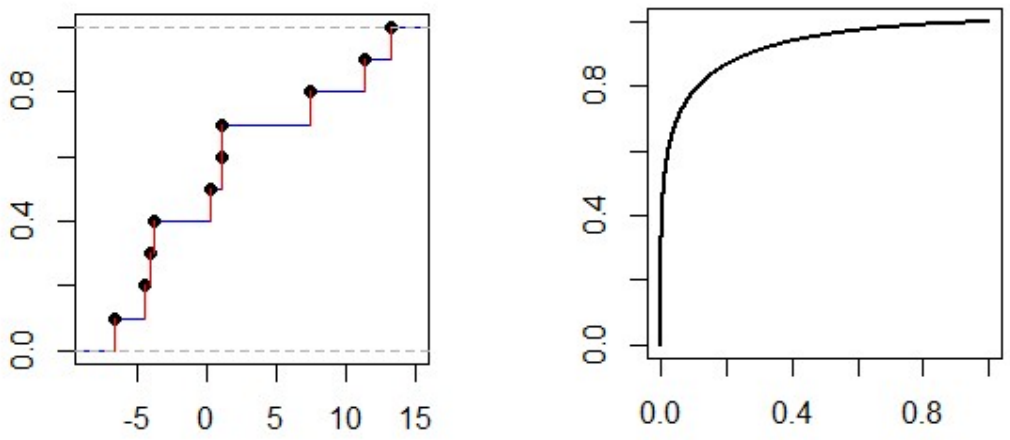
$$SD(E) = \sqrt{\frac{\frac{V_-}{V_- + F_+}\ \frac{F_+}{V_- + F_+}}{V_- + F_+}} = \sqrt{\frac{V_-\ F_+}{(V_- + F_+)^3}}\ .$$

sustituyendo obtenemos que

$$IC(S) = FVP \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{V_+ \, F_-}{(V_+ + F_-)^3}}$$

$$IC(E) = FVP \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{V_- \, F_+}{(V_- + F_+)^3}} \ .$$

Nota. Cabe resaltar que no siempre conocemos la distribución que siguen o bajo qué parámetros están los resultados de la prueba. Esto nos lleva a tener curvas escalonadas (método no paramétrico) o curvas suaves suponiendo la distribución (método paramétrico).
En la Figura 3.6 está representada una curva paramétrica y no paramétrica, es importante elegir según nuestro problema el mejor método para estimar la curva ROC.



Método No Paramétrico. Método Paramétrico.

Figura 3.6: Curvas con diferentes métodos (paramétrico y no paramétrico).

3.1 Métodos no paramétricos

En este método se utiliza la información de los datos directamente y se denomina no paramétrica porque no se hace suposición de las distribuciones de ambos grupos (carece de parámetros).

Método empírico

Si no se hace suposición de las distribuciones, lo más común es sustituirlas por sus funciones de distribución empírica $\hat{f}_{nE}$ y $\hat{f}_{nS}$ de los individuos que presentan el evento y los que no, respectivamente.

Definición 3.2. Dada una muestra aleatoria simple $X_1, ..., X_n$ asociada a la variable aleatoria X con función de distribución F, se define la función de distribución empírica asociada a la muestra como:

$$\hat{F}_n : \mathbb{R} \to [0, 1]$$

$$\text{x}\to \hat{F}_n(x) = \tfrac{1}{n} \sum_{i=1}^{n} \varepsilon_i(x)$$

donde

$$\varepsilon_i(x) = \begin{cases} 1, & \text{si } X_i \leq x \\ 0, & \text{si } X_i > x. \end{cases}$$

Mide la proporción de observaciones menores que un cierto valor fijado.
Las funciones empíricas serán:

$$\hat{F}_{n_E}(x) = \frac{1}{n_E} \sum_{i=1}^{n_E} \varepsilon_i(x)$$

$$\hat{F}_{n_S}(x) = \frac{1}{n_S} \sum_{i=1}^{n_S} \varepsilon_i(x)$$

siendo $X_1, ..., X_{n_E}$ los individuos que presentan el evento y $X_1, ..., X_{n_S}$ los que no.

Definición 3.3. Se define la curva ROC empírica como la curva construida uniendo los $(1 - \hat{F}_{n_S}(x), 1 - \hat{F}_{n_E}(x))$

$$\widehat{ROC}(t) = 1 - \hat{F}_{n_E}(\hat{F}^{-1}_{n_S}(1 - t))$$

siendo n el tamaño de la muestra.
La curva empírica ROC conserva muchas propiedades de la función de distribución empírica, sin embargo el estimador tiene algunos inconvenientes, y puede sufrir una gran variabilidad, particularmente para tamaños de muestra pequeños y comunmente se presentan en la práctica clínica, la curva ROC estimada no es continua y, por lo tanto, su interpretación se vuelve más compleja (Jokiel-Rokita A. (2013)).

Método de la función Kernel

Para superar la falta de suavidad del estimador empírico se utiliza el método de Kernel el cual es una forma de obtener un estimador continuo. Este método también recibe el nombre de estimación suavizada de la curva ROC.

Definición 3.4. Sean $X_{1_E}, ..., X_{n_E}$ los individuos que presentan el evento y sean $X_{1_S}, ..., X_{m_S}$ los que no. Se definen las funciones de densidad Kernel estimadas como

$$\tilde{f}_E(x) = \frac{1}{n_E\ h_1} \sum_{i=1}^{n_E} K_1 \left(\frac{x - x_i}{h_1} \right),$$

$$\tilde{f}_S(x) = \frac{1}{m_S\ h_2} \sum_{i=1}^{n_S} K_2 \left(\frac{x - x_i}{h_2} \right),$$

donde h_1 y h_2 son un par de números positivos llamada ancho de banda estudiadas por (Jokiel-Rokita A. (2013)) y (Pulit M. (2015)) que determina el grado de suavidad de la curva y K_1, K_2 son funciones núcleo.
Las expresiones de las funciones de distribución son:

$$\tilde{F}_E(x) = \frac{1}{n_E} \sum_{i=1}^{n_E} \int_{-\infty}^{x} \frac{1}{h_1} K_1 \left(\frac{u - x_{i_E}}{h_1} \right) du$$

$$\tilde{F}_S(x) = \frac{1}{m_S} \sum_{i=1}^{n_S} \int_{-\infty}^{x} \frac{1}{h_2} K_2 \left(\frac{u - x_{i_S}}{h_2} \right) du.$$

(Zou K. (1998)) propone hacer una estimación gaussiana-kernel como la que sigue

$$\tilde{F}_E(x) = \frac{1}{n_E} \sum_{i=1}^{n_E} \Phi \left(\frac{u - x_{i_E}}{h_1} \right)$$

$$\tilde{F}_S(x) = \frac{1}{m_S} \sum_{i=1}^{m_S} \Phi \left(\frac{u - x_{i_S}}{h_2} \right)$$

siendo Φ la función de densidad de una $N(0, 1)$. Finalmente, bajo estas estimaciones la curva ROC mediante el método Kernel es:

$$\widetilde{ROC}(t) = 1 - \hat{F}_E(\hat{F}_S^{-1}(1 - t)).$$

3.2 Métodos paramétricos

El método paramétrico trata de averiguar la distribución de la variable de decisión de la prueba. Las distribuciones pueden ser diferentes en los grupos de personas y el caso óptimo es que las distribuciones no se solapen es decir que estén lo más separadas posibles. Esto se podría ver como una alta capacidad discriminante y por el contrario si estuviesen solapadas sería una discriminación casi nula.
El modelo más habitual es la distribución normal. Esto es, se supone que el clasificador tanto en la subpoblación con presencia y ausencia de la condición tiene una distribución normal. No obstante, otros modelos también son utilizados para estimar la Curva ROC, como la ditribución logística o la exponencial negativa.

Proposición 3.2. La curva ROC estimada mediante el modelo binormal se determina con dos parámetros, siendo estos:

$$\hat{a} = \frac{\hat{\mu}_S - \hat{\mu}_E}{\hat{\sigma}_E}$$

$$\hat{b} = \frac{\hat{\sigma}_S}{\hat{\sigma}_E}.$$

Con $\hat{\mu}_E$,$\hat{\sigma}_E$, $\hat{\mu}_S$,$\hat{\sigma}_S$ las medias y desviaciones estimadas de enfermos y sanos respectivamente.
Obteniendo la curva determinada por la expresión

$$ROC(t) = 1 - \Phi(\hat{a} + \hat{b} \cdot \Phi^{-1}(1-t)) \quad 0 \leq t \leq 1.$$

Donde Φ representa la función de distribución de la normal estándar.

El modelo binormal se considera comúnmente, y es aplicable cuando los resultados de las pruebas para enfermos como no enfermos siguen distribuciones normales suponiendo que existe una función monótona que transforme a ambos grupos (Faraggi D.(2002)). Si los datos son una transformación como el logaritmo o la raíz cuadrada, hace que los datos sean binormales, entonces los parámetros relevantes pueden estimarse fácilmente por las medias y las variaciones de los valores de prueba en pacientes enfermos y no enfermos.

Después de elegir el modelo, tenemos que estimar los parámetros de dichas distribuciones y la manera más usada es mediante el método de máxima verosimilitud. Una vez de escoger el modelo y estimar los parámetros se puede representar la curva ROC. Esta será suave, no escalonada o si usamos el método empírico, cometeremos mucho error si la selección de distribuciones no es la correcta. Para ellos se hace un juego de hipótesis donde la hipótesis nula sigue la distribución elegida y la hipótesis alternativa es diferente de ella.

La elección del estimador binormal para ajustarse a una curva ROC generalmente se justifica por consideraciones teóricas, o simplemente por conveniencia. Algunos autores también sostienen que el estimador binormal es robusto. La palabra robusto puede tener muchos significados diferentes. Aquí se usa en presencia de una cierta cantidad de observaciones provenientes de una distribución no normal, el estimador binormal producirá resultados confiables. Últimamente, el impacto de la especificación errónea del modelo en los modelos paramétricos o semiparamétricos utilizados en las ciencias de la salud está ganando importancia, ya que los profesionales son conscientes de que los modelos teóricos son solo aproximaciones de la realidad, y los procedimientos estadísticos que dan resultados confiables bajo las desviaciones del modelo son esenciales para resolver problemas reales.

3.3 Métodos semiparamétricos

Por último, otro método de estimación de la Curva ROC es el método semiparamétrico, se basa en agrupar las respuestas en categorias ordenadas y luego se les aplica el método binormal. El término semiparamétrico proviene de mezclar ambos conceptos. Estos métodos asumen la existencia de una transformación de las respuestas del clasificador, aunque desconocida, que permite a ambas distribuciones aproximarse a modelos normales. Entonces sea $ROC = 1 - F_E(F_S^{-1}(1-t))$ la curva ROC asociada a una variable de decisión x con funciones de distribución $F_E(\cdot)$ y $F_S(\cdot)$ en enfermos y sanos. Sea T una transformación monótona y sea $w = T(x)$

$$F_{w,E}(t) = F_E(T^{-1}(t))$$

$$F_{w,S}(t) = F_S(T^{-1}(t))$$

entonces la curva ROC asociada a w es:

$$ROC_w(t) = 1 - F_{w,E}(F_{w,S}^{-1}(1-t))$$

$$= 1 - F_{w,E}(T(F_S^{-1}(1-t)))$$

$$= 1 - F_E(T^{-1}(T(F_S^{-1}(1-t))))$$

$$= 1 - F_E(F_S^{-1}(1-t))$$

$$= ROC(t).$$

Como la transformación de la variable de decisión sigue una distribución normal podemos obtener:

$$ROC(t) = 1 - F_E(F_S^{-1}(1-t))$$

$$= 1 - F_E T T^{-1} F_S^{-1}(1-t)$$

$$= 1 - \Phi\left(\frac{\mu_S + \sigma_S \Phi^{-1}(1-t) - \mu_E}{\sigma_E}\right)$$

siendo μ_S y μ_E las medias de $T(x_s)$ y $T(x_E)$ y sus desviaciones σ_S y σ_E.

Para ilustrar el proceso de construcción de una curva ROC consideremos el ejemplo de (Choi B. (1998)) sobre el diagnóstico de infarto agudo de miocardio usando una concentración de serum creatinine kinase en una muestra de 773 pacientes.

Las respuestas de la prueba diagnóstica son: "definitivamente anormales" (cuando la concentración es superios a 480), "probablemente anormal" (entre 361 y 480), "posiblemente anormal/normal" (entre 241 y 360), "probablemente normal" (entre 121 y 240) y "definitivamente normal" (entre 1 y 120), siendo el verdadero estado de la condición anormal para la presencia infarto agudo de miocardio y normal para la ausencia de infarto agudo de miocardio.

En la siguiente Tabla 3.1 se muestran los resultados de la clasificación:

	>480	361-480	241-360	121-240	1-120	Total
Anormal	9	6	7	6	23	51
Normal	14	12	24	201	471	722

Tabla 3.1: Resultados del diagnóstico de infarto agudo de miocardio.

Para la representación de la curva ROC asociada a esta prueba diagnóstica se tiene en cuenta las distintas clasificaciones para los posibles puntos de corte.

El primero sería con los pacientes con más de 480 son diagnosticados para padecer infarto agudo de miocardio y se obtiene la siguiente Tabla 3.2

	>480	1-480	Total
Anormal	9 (0.176)	42 (0.824)	51
Normal	14 (0.019)	708 (0.981)	722

Tabla 3.2: Matriz de confusión para el criterio de 480.

los valores entre parentesis son sus respectivas fracciones, la *sensibilidad* $FVP = \frac{9}{51} = 0.176$ y la *especificidad* $FVN = \frac{708}{722} = 0.981$, otras medidas del rendimiento de esta clasificación son:

$$AC = 0.928,\ VPP = 0.391,\ VPN = 0.944.$$

Utilizando ahora un diagnóstico menos estricto para la presencia de infarto agudo de miocardio se considera a los pacientes con más de 360 son pronosticados para tener un infarto agudo de miocardio obteniendo la Tabla 3.3

	>360	1-360	Total
Anormal	15 (0.294)	36 (0.706)	51
Normal	26 (0.036)	696 (0.964)	722

Tabla 3.3: Matriz de confusión para el criterio de 360.

los valores entre parentesis son sus respectivas fracciones, la *sensibilidad* $FVP = \frac{15}{51} = 0.294$ y la *especificidad* $FVN = \frac{696}{722} = 0.964$, otras medidas del rendimiento de esta clasificación son:

$$AC = 0.919,\ VPP = 0.366,\ VPN = 0.951.$$

Ahora se considera a los pacientes con más de 240 son pronosticados para tener un infarto agudo de miocardio obteniendo la Tabla 3.4

	>240	1-240	Total
Anormal	22 (0.431)	29 (0.569)	51
Normal	50 (0.069)	672 (0.931)	722

Tabla 3.4: Matriz de confusión para el criterio de 240.

Para esta clasificación se obtienen las estimaciones de la *sensibilidad* $FVP = \frac{22}{51} = 0.431$ y la *especificidad* $FVN = \frac{672}{722} = 0.931$, otras medidas del rendimiento de esta clasificación son:

$$AC = 0.898,\ VPP = 0.306,\ VPN = 0.959.$$

Por último, consideramos el criterio más tolerante para esta prueba diagnóstica, es decir, que sólo los pacientes con 120 o menos no son clasificados para padecer infarto agudo de miocardio, tenemos la siguiente clasificación:

	>120	1-120	Total
Anormal	28 (0.549)	23 (0.451)	51
Normal	251 (0.348)	471 (0.652)	722

Tabla 3.5: Matriz de confusión para el criterio de 120.

Para esta clasificación se obtienen las estimaciones de la *sensibilidad* $FVP = \frac{28}{51} = 0.549$ y la *especificidad* $FVN = \frac{471}{722} = 0.652$, otras medidas del rendimiento de esta clasificación son:

$$AC = 0.646,\ VPP = 0.100,\ VPN = 0.953.$$

Los puntos de corte obtenidos del diagnóstico de infarto agudo de miocardio estan en la siguiente Tabla 3.6

1-E	S
0	0
0.019	0.176
0.036	0.294
0.069	0.431
0.348	0.549
1	1

Tabla 3.6: Puntos de corte.

En la Figura 3.7 se muestra la curva ROC generada a partir de los valores calculados (sensibilidad y el complemento de la especificidad), en cada uno de los puntos de corte del diagnóstico de infarto agudo de miocardio.

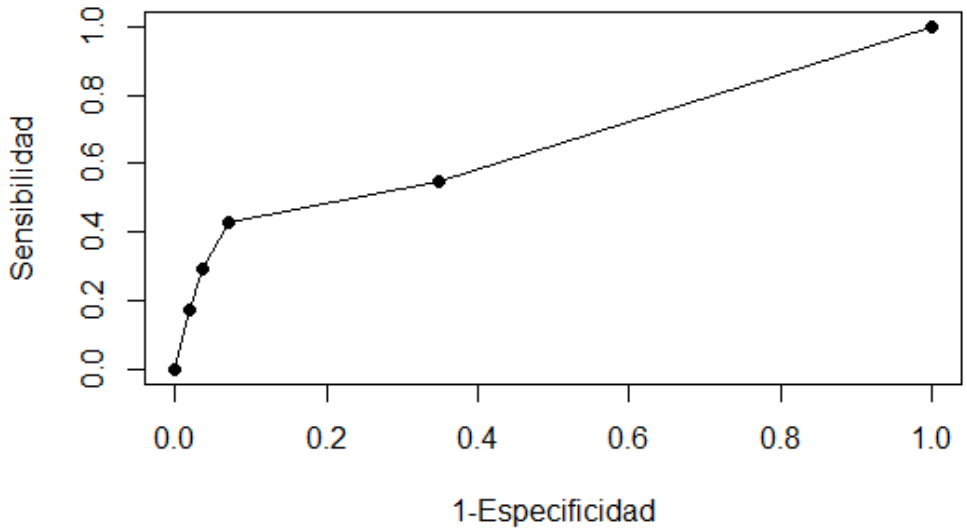


Figura 3.7: Curva ROC de la prueba infarto agudo de miocardio.

Esta representación siempre comienza en (0,0) y termina en (1,1), el punto (0,0) todos los pacientes son clasificados con ausencia de infarto y por tanto, la fracción de verdaderos negativos es uno y la fracción de verdaderos positivos es cero. Sin embargo el punto (1,1) todos los pacientes son clasificados con presencia de infarto, obteniendo la fracción de verdaderos positivos es uno y la fracción de verdaderos negativos es cero.

En general, una curva ROC generada a partir de un conjunto finito de respuestas es una función escalonada o lineal a trozos como la curva ROC en la Figura 3.7, que se aproxima a la verdadera curva ROC cuando el número de casos tiende a infinito.

Además, se puede obtener una curva ROC por debajo de la diagonal, cuando esto sucede estos rendimientos pueden ser mejorados considerando el cambio de las respuestas del clasificador (inversión de las predicciones) que se corresponde con una simetría de la curva ROC.

La curva ROC es necesariamente monótona no decreciente, como consecuencia de la relación existente entre la sensibilidad y la especificidad de un clasificador, sin embargo, no necesariamente tiene por qué ser convexa (Lloyd C. (2001)).

Capítulo 4

Medidas para un clasificador

Después de dibujar nuestra curva ROC necesitamos una medida con la cual interpretar si la variable de estudio discrimina bien a la muestra o en caso contrario no lo hace.

4.1 Exactitud de un clasificador

Definición 4.1. Se define la exactitud o acracidad, accuracy (AC) de una variable de decisión como la probabilidad de discriminar correctamente, en otras palabras se refiere a lo cerca que está el resultado de una medición del valor verdadero.
Entonces en términos probabilísticos tendremos:

$$AC = P(y = 1|D = 1)\ P(D = 1) + P(y = 0|D = 0)\ P(D = 0)$$

$$AC = S\ prevalencia + E\ (1 - prevalencia)$$

es decir, una combinación de la sensibilidad y la especificidad ponderadas por la prevalencia y su complemento respectivamente, entonces dada una muestra se tiene un estimador basado en la sensibilidad, especificidad y prevalencia es:

$$\widehat{AC} = \frac{V_+ + V_-}{V_+ + V_- + F_+ + F_-} = \frac{resultados\ acertados}{total\ de\ la\ muestra}.$$

Este índice es calculado a través de la Tabla 2.1 como la proporción de verdaderos positivos y negativos en la muestra.

Definición 4.2. La precisión se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas. Cuanto menos es la dispersión mayor es la precisión. Se representa por la proporción entre el número de predicciones correctas y el total de predicciones.

Así un clasificador ideal o perfecto es aquel cuyas respuestas son correctas, y por tanto su exactitud es uno, $\widehat{AC} = 1$, dado que $F_+ = F_- = 0$, esto es, no produce clasificacioes erróneas.

De la fórmula anterior se deduce la predominancia (PD);

$$PD = \frac{\widehat{AC} - E}{S - E} = \frac{TCP}{V_+ + V_-}$$

La predominancia se define como el porcentaje de individuos con presencia de la condición en la muestra, es decir, el estimador de la prevalencia.

4.2 Índice de Youden

El índice de Youden esta definido por la diferencia de las proporciones de respuestas positivas correctas e incorrectas, es decir la sensibilidad y especificidad menos uno. Este refleja la diferencia entre los verdaderos positivos y los falsos positivos. Un buen test debe tener una alta especificidad y una alta sensibilidad.
El índice de Youden fue introducido por W. J. Youden en 1950 y se define como (Estrada J. (2015)):

$$\begin{aligned} J &= 1 - [P(y = 1|D = 0) + P(y = 0|D = 1)] \\ &= 1 - [1 - E + 1 - S] \\ &= S + E - 1 \\ &= S + E - 1 \\ &= FVP + (1 - FFP) - 1 \\ &= FVP - FFP. \end{aligned}$$

Esto es, una medida basada en la sensibilidad y especificidad que no depende de la prevalencia de la condición. Si la sensibilidad y la especificidad son menores que 0.5 el índice J tomará valores menores a 1 lo que diría que tiene un comportamiento negativo con la enfermedad, y no es tan común en la práctica. Mientras que si toman valores iguales a 0.5 para sensibilidad y especifidad el índice J toma valor cero. Por último si el valor de la sensibilidad y especificidad toman valores cercanos a 1 como máximo, el valor del índice J tomaría valor máximo como 1 indicando un test excelente.
Bajo esto tenemos que el índice toma valores en [0,1] si toma el valor 1 es un test perfecto y 0 el que no nos da información para detectar la enfermedad. Características del índice de Youden:

• El índice tiene valor cero cuando tiene la misma proporción de positivos en el grupo de enfermos y sanos.
• El índice se vuelve 1 únicamente cuando los falsos positivos y falsos negativos no están presentes. Si solo se presenta un tipo de error, el índice es controlado por ese error.
• Es posible calcular un error estándar para el índice, siendo el tamaño de la muestra un factor que influye naturalmente la confiabilidad en la estimación.

El índice de Youden es un indicador con ventajas siempre que tenga la misma importancia la sensibilidad y la especificidad.

4.3 Razones de verosimilitud

La capacidad discriminante del clasificador puede ser expresada en función de la razón o taza de verosimilitud (LR).

Definición 4.3. Se define la razón de verosimilitud (LR) como el cociente de la probabilidad de respuestas positivas o negativas bajo presencia del evento entre la probabilidad de respuesta positiva o negativa bajo ausencia del evento de interés.

Razón de verosimilitud postiva (LRP):

$$LRP = \frac{P(Y=1|D=1)}{P(Y=1|D=0)},$$

$$LRP = \frac{S}{1-E} \ .$$

Razón de verosimilitud negativa (LRN):

$$LRN = \frac{P(Y=0|D=1)}{P(Y=0|D=0)},$$

$$LRN = \frac{1-S}{E} \ .$$

Los valores de LRP y LRN recorren el intervalo $[0, \infty)$. Un valor mayor que 1 significa que hay más respuestas positivas en individuos que presentan el evento que en los que no, y viceversa para un valor menor que 1.
Estimación de la razón de verosimilitud positiva:

$$\widehat{LRP} = \frac{FVP}{FFP} \ .$$

Estimador de la razón de verosimilitud negativa:

$$\widehat{LRN} = \frac{FFN}{FVN} \ .$$

4.4 Odds ratio

Otra medida de la exactitud de un clasificador se basa en las *odds* (ventajas o preferencias) de las respuestas.

Definición 4.4. Se define el *odds* de un suceso como el cociente entre su probabilidad y la de su complemento:

$$Odds(suceso) = \frac{P(suceso)}{1 - P(suceso)}.$$

Así, la ventaja de un suceso puede interpretarse como sigue:

• $Odds(suceso) = \frac{P(suceso)}{1 - P(suceso)} > 1$, entonces, la probabilidad de ocurrencia del suceso es mayor que la no ocurrencia.

• $Odds(suceso) = \frac{P(suceso)}{1 - P(suceso)} < 1$, entonces, la probabilidad de ocurrencia del suceso es menor que la no ocurrencia.

• $Odds(suceso) = \frac{P(suceso)}{1 - P(suceso)} = 1$, entonces, la probabilidad de ocurrencia del suceso y la de su complemento son igual.

Además, la probabilidad del suceso puede expresarse en función de su *odds*

$$P(suceso) = \frac{Odds(suceso)}{1 + Odds(suceso)}.$$

En este sentido, dos tipos de *odds* son de interés para analizar el rendimiento de un clasificador

$$Odds_{presencia} = \frac{P(Y = 1|D = 1)}{P(Y = 0|D = 1)},$$

es decir, la ventaja del clasificador para una respuesta positiva frente a una negativa bajo la presencia de la condición, y

$$Odds_{ausencia} = \frac{P(Y = 1|D = 0)}{P(Y = 0|D = 0)},$$

es decir, la ventaja del clasificador para una respuesta positiva frente a una negativa bajo la ausencia de la condición.

Escrito de otra manera tenemos

$$Odds_{presencia} = \frac{S}{1 - S}$$

y

$$Odds_{ausencia} = \frac{1 - E}{E}$$

apartir de estas dos tasas se define el *odds ratio*.

Definición 4.5. Se define el *odds ratio* como el siguiente cociente de *odds*:

$$Odds\ ratio = \frac{Odds_{ausencia}}{Odds_{presencia}},$$

siendo su estimador

$$Odds\ ratio = \frac{S\ E}{(1-S)\ (1-E)} = \frac{V_+\ V_-}{F_-\ F_+}.$$

Según sus valores indican:

• $Odds\ ratio > 1$ tenemos mayor ocurrencia de respuesta positiva cuando el evento de interés está presente.

• $Odds\ ratio < 1$ tenemos menor ocurrencia de respuesta positiva cuando el evento de interés está presente.

• $Odds\ ratio = 1$ igual ocurrencia en ambos casos, es decir, $S = 1 - E$.

Nota. Si alguno de los elementos de la matriz de confusión es nulo, entonces no puede estimarse el *odds ratio*.

Nota. El *odds ratio* está totalmente determinado por la sensibilidad y especificidad, y por tanto, no dependen de la prevalencia de la condición.

4.5 Índice de discriminación

Definición 4.6. El índice de discriminación es una medida para cuantificar que tan separadas están graficamente las funciones de densidad del grupo de individuos que presenta el evento y los que no. Su expresión viene dada por:

$$\delta = \frac{separacion}{dispersion} = \frac{|\mu_E - \mu_S|}{\sigma}.$$

Siendo μ_E y μ_S las medias de nuestra variable de decisión en el grupo que presenta el evento y en el que no, por otro lado σ representa el error estándar de la variable de decisión en la muestra total.
Este índice mide la proporción de pacientes correctamente diagnosticados, pero trata por igual a positivos y negativos, verdaderos o falsos.

4.6 Área bajo la curva

El Área Bajo la Curva ROC (Area Under the Curve) estima la capacidad de distinguir o de "discriminar" entre enfermos y no enfermos que tiene una prueba diagnóstica.

Definición 4.7. Sea $ROC(t)$ la función asociada a la curva ROC. Se define el área bajo la curva ROC como:

$$AUC = \int_0^1 ROC(t)\, dt.$$

El rango de valores va de 0.5 a 1 donde 1 es cuando están perfectamente diferenciado los grupos y .5 cuando no hay capacidad discriminante.
• Baja exactitud $[0.5, 0.7)$.
• Se utiliza para algunos propósitos $[0.7, 0.9)$.
• Exactitud alta $[0.9, 1]$.

El área bajo la curva ROC tiene una interpretación interesante. Según (Franco M. (2007)) Bamber (1975) y Hanley and McNeil (1982), se define como la probabilidad de clasificar correctamente un par de individuos uno sano y otro enfermo, son seleccionados al azar de la población.

Por último, el área bajo la curva ROC también puede interpretarse como un promedio de la sensibilidad para todos los posibles valores de especificidad. Y análogamente, como promedio de la especificidad para todos los posibles valores de sensibilidad.

4.7 Cálculo del área bajo la curva

4.7.1 Método no paramétrico

Consideramos métodos de estimación del área bajo la curva ROC sin asumir ninguna distribución del clasificador sobre las dos poblaciones dadas la cual es la que presenta la enfermedad y la que no.

Regla Trapezoidal

Si utilizamos el método empírico para la construcción de la curva ROC podemos calcular el área con el método del trapecio teniendo como resultado una curva de la forma de escalera. La fórmula está dada por:

$$AUC = \sum_{t=1}^{T} \frac{1}{2}(FFP_t - FFP_{t-1})\ (FVP_t + FVP_{t-1}),$$

siendo (FFP_t, FVP_t) las fracciones de falsos positivos y verdaderos positivos calculadas para $t = 0.1..., T$ es el conjunto de T puntos de la Curva ROC.

Método de la función Kernel

Para la curvas ROC no paramétricas suavizadas está la función kernel y podemos calcular el área como:

$$AUC = \int_0^1 ROC(t)\, dt$$

y

$$\widehat{ROC}(t) = 1 - \hat{F}_E(\hat{F}_S^{-1}(1-t))$$

de modo que obtendremos

$$AUC = \int_0^1 \widehat{ROC}(t)\, dt = \int_0^1 (1 - \hat{F}_E(\hat{F}_S^{-1}(1-t)))\, dt.$$

4.7.2 Métodos paramétricos y semiparamétricos

Al igual que en la estimación de la curva ROC de un clasificador, asumiendo un modelo de distribución particular del clasificador en cada subpoblación determinada por el estado de la condición, se estima de forma paramétrica el área bajo la Curva ROC, aunque cualquier método va a estar sujeta a un error. La forma de cuantificar un error es mediante un intervalo de confianza. Para ello tenemos que fijar un punto de corte y calcular un intervalo de confianza para la sensibilidad y especificidad.
Una vez asegurada la distribución y estimadas $F_E(x)$ y $F_S(x)$ el cálculo del área será el mismo:

$$AUC = \int_0^1 (1 - \hat{F}_E(\hat{F}_S^{-1}(1-t)))\, dt.$$

También para el caso binormal queda determinada por los parámetros a y b obtenemos:

$$\widehat{AUC} = \int_0^1 \widehat{ROC}(t)\, dt = \int_0^1 (1 - \Phi(\hat{a} + \hat{b} \cdot \Phi^{-1}(1-t)))\, dt.$$

Siendo Φ la función de densidad de una distribución normal estándar (Faraggi D. (2002)).

Definición 4.8. El área parcial de la curva ROC esta definida como:

$$AUC_{(a,b)} = \int_a^b ROC(t)\, dt.$$

Nota. Dos curvas ROC son iguales si tienen la misma área pero dos áreas iguales no tienen por que ser las mismas curvas ROC.

El análisis de un clasificador, que se maneja en este trabajo corresponde a un clasificador con respuestas dicotómicas, lo que dificultaría aquellos casos de variables ordinales o continuas. Se pueden dicotomizar mediante la elección de decisión o punto de corte.

El punto de corte no será único en estos casos, además cada punto de corte le corresponde un par de valores (1-E,S) se obtiene un conjunto de pares para analizar el rendimiento de un clasificador.

Capítulo 5

Punto de corte

En general se han destacado la sensibilidad y la especificidad, además la representación de la curva ROC está determinada por ambos valores asociados a cada posible punto de corte del clasificador, reflejando la relación entre ambas medidas, dado que la variación del punto de corte conlleva un cambio de sensibilidad y especificidad, implicando el incremento de una y la reducción de la otra. Por tanto las curvas ROC pueden ser utilizadas para encontrar un punto de corte óptimo.

Se podría pensar que el mejor punto de corte es el punto (0,1), sin embargo es importante controlar cómo afectan las falsas alarmas y los fracasos en la selección del punto de corte óptimo, es decir, los costes asociados a las clasificaciones incorrectas.

(Altman D. (1998)) dice que si el coste de un falso positivo y un falso negativo no son iguales, entonces el punto de corte (0,1) no necesariamente es el punto de corte óptimo.

Por lo tanto la calidad de un clasificador se considera más en términos de minimización de costes que de errores, entonces el punto de corte óptimo debe elegirse de manera que el clasificador presente una relación óptima entre los costes de fallo y fracasos, es decir, la mejor compesación entre el coste de fallos al detectar positivos (falsas alarmas) frente al coste de incrementar los fracasos al descartar la presencia (falsos negativos).

Después de haber hecho la curva ROC y ver que el área bajo la curva es apropiada o tiene un alto nivel discriminante, nos queda escoger el punto de corte o valor umbral. Podemos escoger el punto de corte más alto con el índice de Youden (hay un detalle que estaríamos escogiendo la sensibilidad y especificidad de manera conjunta). Es por eso que a veces requerimos otros métodos porque queremos nuestra prueba muy sensible o viceversa. Tenemos dos caminos, escoger el punto de mínima distancia al vértice (0,1) o minimizar los costes de los resultados erróneos.

Lema 5.1. Dada una curva ROC el punto de corte correspondiente al par $(1-E, S)$ más cercano al $(0, 1)$ es aquel cuya recta tangente tiene pendiente

$$m = \frac{p\ P(falsos\ positivos)}{(1-p)\ P(falsos\ negativos)},$$

siendo p la prevalencia del evento en la población.

Demostración.
Sea $y = f(x)$ la función que define la curva ROC con $(x, y) = (1-E, S)$. La distancia de un punto cualquiera al vértice $(0, 1)$ es una función que puede ser expresada como sigue

$$dis(x) = \sqrt{(x-0)^2 + (y-1)^2} = \sqrt{x^2 + (f(x)-1)^2}.$$

Buscamos el mínimo

$$\frac{1}{2\sqrt{x^2 + (f(x)-1)^2}}(2x + 2(f(x)-1)f'(x)) = 0$$

$$2x + 2(f(x)-1)f'(x) = 0$$

$$2(f(x)-1)f'(x) = -2x$$

$$f'(x) = \frac{-2x}{2(f(x)-1)}$$

$$f'(x) = \frac{-x}{f(x)-1}$$

$$f'(x) = \frac{x}{1-f(x)}$$

es decir el punto será:

$$\frac{1-E}{1-S} = \frac{1-P(Y=0|D=0)}{1-P(Y=1|D=1)} = \frac{P(Y=1|D=0)}{P(Y=0|D=1)},$$

que será igual a:

$$\frac{P(\{Y=1\}\cap\{D=0\})}{P(D=0)}\ \frac{P(D=1)}{P(\{Y=0\}\cap\{D=1\})} = \frac{p\ P(falsos\ postivos)}{(1-p)\ P(falsos\ negativos)}\ .$$

Por tanto el mínimo se alcanzará en el punto cuya pendiente sea m

$$m = \frac{p\ P(falsos\ positivos)}{(1-p)\ P(falsos\ negativos)}\ .$$

Lema 5.2. Sea una curva ROC asociada a un diagnóstico sobre un evento y considérense los costes previamente fijados, de los resultados erróneos, entonces el punto de corte correspondiente al par $(1-E, S)$ que minimiza dichos costes es aquel cuya recta tangente tiene por pendiente

$$m = \frac{costes\ falsos\ positivos\ (1-p)}{costes\ falsos\ negativos\ p},$$

siendo p la prevalencia del evento.

Demostración.
Se define el coste medio esperado como

$$C_{esp} = C_0 + C_{V_+}\ P(V_+) + C_{V_-}\ P(V_-) + C_{F_+}\ P(F_+) + C_{F_-}\ P(F_-)$$

siendo

$$\begin{cases} C_0, & \text{coste base} \\ C_{V_+}, & \text{coste de verdaderos positivos} \\ C_{F_+}, & \text{coste de falsos positivos} \\ C_{F_-}, & \text{coste de falsos negativos} \\ C_{V_-}, & \text{coste de verdaderos negativos,} \end{cases}$$

también tenemos

$$\begin{cases} P(V_+) = P(Enfermo)\ P(y=1 \mid Enfermo) = p\ S \\ P(V_-) = P(Sano)\ P(y=0 \mid Sano) = (1-p)\ E \\ P(F_+) = P(Sano)\ P(y=0 \mid Sano) = (1-p)\ (1-E) \\ P(F_+) = P(Enfermo)\ P(y=0 \mid Sano) = p\ (1-S), \end{cases}$$

sustituyendo

$$C_{esp} = C_0 + C_{V_+}\ p\ S + C_{V_-}\ (1-p)\ E + C_{F_+}\ (1-p)\ (1-E) + C_{F_-}\ p\ (1-S)$$

otra forma sería

$$C_{esp} = C_0 + C_{V_+}\ p\ f(x) + C_{V_-}\ (1-p)\ (1-x) + C_{F_+}\ (1-p)\ x + C_{F_-}\ p\ (1-f(x))$$

reagrupando tenemos:

$$C_{esp} = C_0 + C_{V_-}\ (1-p) + C_{F_-}\ p + p\ (+C_{V_+} - C_{F_-})\ f(x) + (1-p)\ (C_{F_+} - C_{V_-})\ x.$$

Buscamos el mínimo de ésta función, derivando e igualando a 0.

$$p\ (C_{V_+} - C_{F_-})\ f'(x) + (1-p)\ (C_{F_+} - C_{V_-}) = 0$$

despejando obtenemos:

$$f'(x) = \frac{(C_{F_+} - C_{V_-})\ (1-p)}{-(+C_{V_+} - C_{F_-})\ p}$$

si consideramos que los costes de los resultados acertados son nulos, es decir, $C_{V_-} = 0 = C_{V_+}$ nos queda:

$$f'(x) = \frac{C_{F_+}\ (1-p)}{C_{F_-}\ p}$$

el mínimo se alcanzará en el punto que tenga por pendiente m

$$m = \frac{C_{F_+}\ (1-p)}{C_{F_-}\ p}.$$

Otro método para la selección del punto de corte óptimo es mediante coste medio total del clasificador.
La matriz de confusión tiene asociada una matriz de costes se muestra en la Tabla 5.1

	Positiva	Negativa
Presencia	*Éxito* (C_{VP})	*Fracaso* (C_{FN})
Ausencia	*Falsa alarma* (C_{FP})	*Rechazo correcto* (C_{VN})

Tabla 5.1: Matriz de costes.

Así, el coste medio total del rendimiento de un clasificador, viene dado en función de los costes que conlleva cada uno de los posibles resultados de la clasificación ponderados por sus correspondientes probabilidades de ocurrencia, es decir,

$$c_t = C_0 + P(VP)\ C_{VP} + P(FN)\ C_{FN} + P(FP)\ C_{FP} + P(VN)\ C_{VN}$$

donde C_0 denota aquellos costes del rendimiento de un clasificador debidos a causas diferentes a las respuestas del clasificador.

Además las ponderaciones asociadas a estos costes, están determinadas por:

$$P(VP) = P(D=1)\ P(y=1 \mid D=1)$$

$$P(FP) = P(D = 0)\ P(y = 1 \mid D = 0)$$

$$P(FN) = P(D = 1)\ P(y = 0 \mid D = 1)$$

$$P(VN) = P(D = 0)\ P(y = 0 \mid D = 0)$$

tenemos

$$S = P(y = 1 \mid D = 1),\ E = P(y = 0 \mid D = 0) \text{ y } PD = P(D = 1)$$

se tiene que

$$P(VP) = PD\ S, \quad P(FP) = (1 - PD)\ (1 - E) = (1 - PD)\ FFP$$

$$P(FN) = PD\ (1 - S), \quad P(VN) = (1 - PD)\ E = (1 - PD)\ (1 - FFP)$$

por lo que el coste medio total de un clasificador depende de la sensibilidad, especificidad y las consecuencias de las decisiones del clasificador, asi como de la prevalencia de la condición.

Esta relación del coste medio total incluye el modelo dado en (Pepe M. (2003)) para situaciones de diagnóstico médico:

$$c_t = C_0 + P(VP)\ C_{VP} + P(FN)\ C_{FN} + P(FP)\ C_{FP}$$

dado que en una prueba diagnóstica, una respuesta positiva lleva al menos el coste del propio tratamiento de la enfermedad, tanto en el caso de diagnóstico correcto (C_{VP}) como incorrecto (C_{FP}) y una respuesta negativa sobre un paciente con la enfermedad retrasa su adecuado tratamineto, por lo que requerirá un posterior tratamiento más severo, en generalcon mayor coste C_{FN}. Sin embargo, los pacientes con respuestas negativas diagnosticados correctamente, no se verán sometidos a tratamiento alguno, $C_{VN} = 0$.

El punto de la curva ROC con mínimo valor de:

$$c_t = C_0 + S\ P(D = 1)\ (C_{VP} - C_{FN}) + FFP\ P(D = 0)\ (C_{FP} - C_{VN}) + P(D = 0)\ C_{VN} + P(D = 1)\ C_{FN}$$

donde cada punto de corte tiene asociado un punto (FFP,S) de la curva.

Ahora el punto de corte óptimo correspondiente al punto de la curva ROC con pendiente óptima, (Franco M. (2007)).

$$m = \frac{P(D = 0)}{P(D = 1)}\ \frac{C_{FP} - C_{VN}}{C_{FN} - C_{VP}}$$

la pendiente óptima está determinada por la prevalencia de la condición y la matriz de costes asociados a los distintos resultados de la clasificación, siendo su estimador:

$$m = \frac{1 - PD}{PD}\ \frac{C_{FP} - C_{VN}}{C_{FN} - C_{VP}}.$$

Capítulo 6

Software estadístico para la curva ROC

En este capítulo se aplican algunos métodos a un estudio de 286 pacientes con cáncer de mama que no habían recibido ninguna terapia, se obtuvieron del banco de tumores en el Centro Médico Erasmus (Rotterdam, Países Bajos), la edad media de los pacientes fue de 53 años (rango 26-83 años) el seguimiento de los pacientes fue de 5 años. Los datos utilizados en este trabajo están disponibles públicamente en (Domrachev M. (2002)) y (Wang Y. (2005)). Se recopilaron datos de expresión génica y se midieron sus variables fenotípicas, el estado de ER, es el evento de recaída cerebral y cáncer de mama. En este trabajo, la expresión génica del receptor de progesterona (PgR) se usa como un biomarcador para predecir el estado de ER de los pacientes (Woo S. (2015)).

Antes de aplicar cualquier método es necesario hacer un análisis para entender mejor cuales son las características de las variables involucradas. Se tiene que de la base de 286 registros los datos son completos, por lo que no se pierde información, hay 209 individuos que presentan el evento o sea cáncer de mama y recaída cerebral y 77 fueron favorables. De las 2 variables se tiene que una es continua (PgR) y otra es categórica (ER).
La librería usada para la aplicación de las curvas ROC es el denominado pROC Display and Analyze ROC curves (Robin X. (2011)) y (Quintela A. (2012)), permite representar curvas ROC, calcular el área bajo ellas, calcular intervalos de confianza. Las primeras líneas de nuestro código serán:

```
install.packages "pROC"
library(pROC)
```

Después tenemos que leer los datos que fueron guardados en el archivo dato.txt, para ello usamos la función read.table():

$$\text{datos} < - \text{read.table(file} = \text{"datos.txt"}, \text{header} = \text{TRUE}$$

Con las funciones summary() y str() se obtiene el resumen de los datos:

Summary (datos)			
PgR		ER	
Min.	: 140.500	Min.	: 0.0000
1st Qu.	: 726.900	1st Qu.	: 0.0000
Median	: 6865.70	Median	: 1.0000
3rd Qu.	: 12572.4	3rd Qu.	: 1.0000
Max.	: 29522.0	Max.	: 1.0000

str (datos))
data.frame: 286 obs. of 2 variables:
$ PgR: num 14609 13657 6886 307 472 ...
$ ER: int 1 1 1 0 1 1 1 1 1 1 ...

names(datos)
"PgR" "ER"

Para representar las curvas ROC necesitaremos a cada variable por separado.

PgR <− datos [, 1]
ER <− datos [, 2]

Tenemos los datos que representa la curva ROC. Cuyos elementos más importantes son:
• La variable de la condición real ER.
• La variable que se pretende estudiar PgR.
• Controls, cases: Se introducen las variables PgR y ER como dos vectores donde una es el marcador y la otra el estado del individuo.
• Percent=true: Los resultados son expresados en fracción.
• Na.rm=true: Elimina a los individuos que no tengan datos completos.
• Direction: Cambia el grupo de casos por el de controles, es decir, este argumento nos sirve para cuando el número de enfermos es menor en la prueba que los sanos y necesitemos invertir la positividad de la curva para que salga cóncava y no convexa.

• Smooth=true: Suavización de la curva.
• Auc=true: Se calcula el área bajo la curva.
• Ci=true: Es el intervalo de confianza (Auc).
• Plot=true: Gráfica de la curva.
• Smooth.method: Específica las densidades que vamos ajustar en ambos grupos (enfermos y sanos).
Obtendremos la Curva ROC no paramétrica de la variable PgR como se muestra:

```
ROCER <- roc(ER, PgR, percent = FALSE, na.rm = TRUE, direction =
c("auto","<",">"), smooth = FALSE, auc = TRUE, ci = TRUE, plot = TRUE)
ROCER
```

Nos devuelve:

```
Data: PgR in 77 controls (ER 0) < 209 cases (ER 1).
Area under the curve: 0.9393
95% CI: 0.9076 - 0.9711 (DeLong)
```

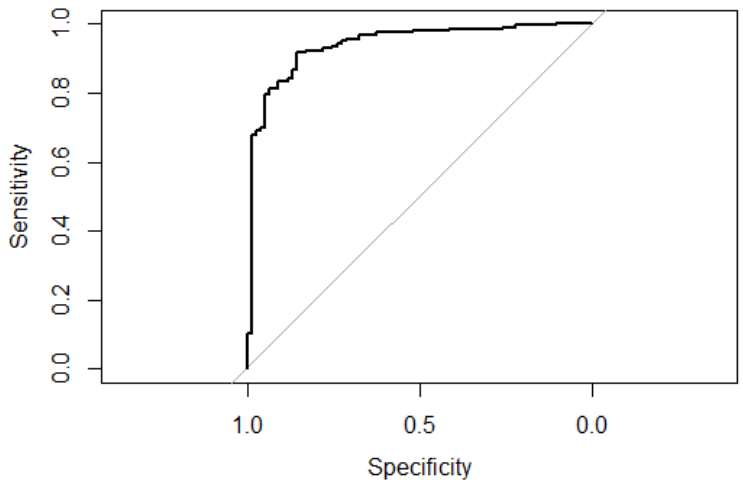


Figura 6.1: Curva ROC no paramétrica de la variable PgR.

Cambiando el argumento Smooth obtenemos la curva ROC paramétrica de la variable PgR donde las distribuciones de sanos y enfermos han sido ajustadas a una distribución binormal. También podemos ajustarla a una suavización Kernel usando smooth.method = c("density"). Además, eliminando el argumento ci.method = NULL calcula el intervalo de confianza para el área realizando 2000 muestras bootstrap.

```
ROCERSmooth <- roc(ER, PgR, percent = FALSE, na.rm = TRUE, direction
= c("auto","<",">"), smooth = FALSE, auc = TRUE, ci = TRUE, plot =
TRUE), smooth.method = "binormal", desity = NULL)
ROCERSmooth
```

Nos devuelve:

```
Data: PgR in 77 controls (ER 0) < 209 cases (ER 1).
Smoothing: binormal
Area under the curve: 0.9283
95% CI: 0.8901- 0.9642
```

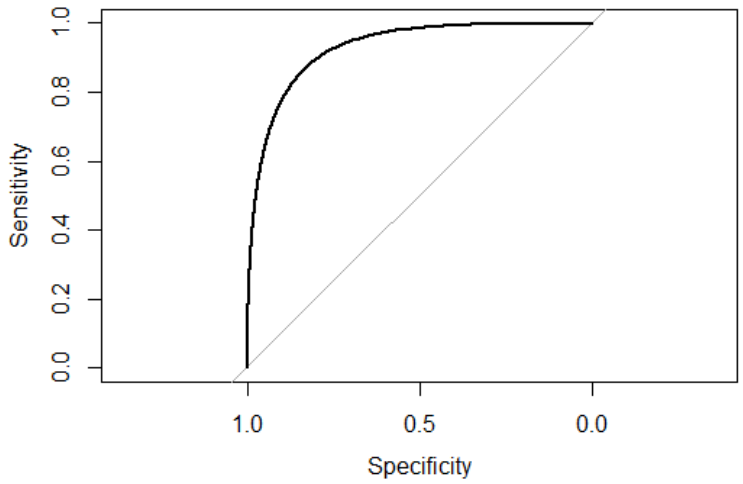


Figura 6.2: Curva ROC paramétrica de la variable PgR.

Por otra parte, también podemos obtener directamente el área mediante la función auc(), su argumento es el nombre de la curva. Al ejecutarlo en nuestro script observamos que las estimaciones del área por los método no paramétrico y paramétrico son parecidas.

```
auc(ROCER)
Area under the curve: 0.9393
auc(ROCERSmooth)
Area under the curve: 0.9283
```

Con la función Smooth() podemos obtener la suavización de nuestra curva mediante los métodos: binormal, density (el correspondiente al método Kernel), ftdistr. Pueden darse considerables cambios en las estimaciones del área según el método utilizado, esto dependerá de la precisión de las estimaciones y tamaños muestrales.

```
smooth(ROCER, method = c("binormal"))

Call:
smooth.roc(roc = ROCER, method = c("binormal"))

Data: PgR in 77 controls (ER 0) < 209 cases (ER 1).
Smoothing: binormal
Area under the curve: 0.9283
```

```
smooth(ROCER, method = c("density"), n=512, bw="nrd0")

Call:
smooth.roc(roc = ROCER, method = c("density"), n = 512, bw = "nrd0")

Data: PgR in 77 controls (ER 0) < 209 cases (ER 1).
Smoothing: density (bandwidth: nrd0; adjust: 1)
Area under the curve: 0.8939
```

```
smooth(ROCER, method = c("fitdistr"), n=512, bw="nrd0")

Call:
smooth.roc(roc = ROCER, method = c("fitdistr"), n = 512, bw = "nrd0")

Data: PgR in 77 controls (ER 0) < 209 cases (ER 1).
Smoothing: fitdistr
Area under the curve: 0.9285
```

Para obtener directamente el intervalo de confianza del área de la curva podemos usar las funciones ci() y ci.auc(). A ambas se les puede añadir como argumento el nombre del elemento ROC que hayamos creado o los datos en dos vectores, uno correspondiente a sanos y otro a enfermos. Además, con el argumento method se le puede especificar si se quiere calcular con el estadístico DeLong o mediante el método Bootstrap.

```
ci(ROCER)
95% CI: 0.9076 - 0.9711 (DeLong)
ci.auc(ROCER)
95% CI: 0.9076 - 0.9711 (DeLong)
```

Otras funciones de interés son ci.sp() y ci.se(), que permiten calcular el intervalo de confianza de la especificidad y sensibilidad para cada valor estimado de sensibilidad y especificidad respectivamente. Sus argumentos son similares a las funciones anteriores, si no se le especifica lo contrario realizará 2000 muestras bootstrap. En la salida obtenemos una tabla de cuatro columnas, la primera corresponde al recorrido de valores de sensibilidad o especificidad yendo desde 0 hasta 1 sumando 0,1, la segunda y la cuarta corresponden a los extremos de los intervalos y la tercera una media de estos.

ci.se(ROCER)

95% CI (2000 stratified bootstrap replicates):

sp	se.low	se. median	se. high
0.0	1.00000	1.0000	1.0000
0.1	0.98560	1.0000	1.0000
0.2	0.97610	0.9952	1.0000
0.3	0.96650	0.9856	1.0000
0.4	0.96170	0.9856	1.0000
0.5	0.95690	0.9809	0.9952
0.6	0.94260	0.9757	0.9952
0.7	0.90910	0.9569	0.9856
0.8	0.84210	0.9282	0.9665
0.9	0.70800	0.8325	0.9330
1.0	0.06699	0.1196	0.7560

ci.sp(ROCER)

95% CI (2000 stratified bootstrap replicates):

se	sp.low	sp. median	sp. high
0.0	1.00000	1.0000	1.0000
0.1	0.96100	1.0000	1.0000
0.2	0.96100	0.9870	1.0000
0.3	0.96100	0.9870	1.0000
0.4	0.96100	0.9870	1.0000
0.5	0.96100	0.9870	1.0000
0.6	0.96100	0.9870	1.0000
0.7	0.90910	0.9610	1.0000
0.8	0.84420	0.9351	0.9870
0.9	0.71430	0.8571	0.9351
1.0	0.03896	0.1299	0.4545

El software estadístico R, permite la implementación de diversos paquetes de forma libre. Ese es el caso del paquete publicado en el 2014 denominado OptimalCutpoint (López M. (2014)) y (Caro A. (2017)) el cual tiene funciones para el análisis de selección de puntos de corte óptimos en un test binario.

Las principales funciones de este paquete son:
Optimal.cutpoints: Es el valor umbral óptimo.
Control.cutpoints: función que genera distintos parámetros para el control de la función optimal.cutpoints.
La función principal tiene los siguientes argumentos:
• Optimal.cutpoints(X, status, tag.healthy, methods, data, direction= c("<", ">"), pop.prev = NULL, ci.fit = FALSE, conf.level = 0.95,).
• X: Es la que contiene el resultado de un test diagnóstico.
• Status: Es donde se guarda la información para distinguir enfermos de sanos.
• Tag.healthy: Es un argumento que identifica el valor con el cual esta codificado los individuos sanos en la variable status.
• Methods: Se conoce el método para la selección del valor umbral óptimo.
• Data: Se debe tener las siguientes variables: el valor de la prueba diagnóstica, y la variable de estado.
• Direction: Es la dirección en la cual la Curva ROC es calculada. Se ponen implícito a los individuos sanos con menor valor, y a los individuos con un valor mayor son los enfermos. Si el caso es contrario, el argumento direction, debe ser especificado con ">" .
• Pop.prev: Es para el valor de la prevalencia, se toma en cuenta el número de personas en la muestra.
• Ci.fit: Se calculan los intervalos de confianza para difernetes métodos.
• Conf.level: Nivel de confianza en las estimaciones, (0.95).
Generemos el punto de corte óptimo

```
cortehta <- optimal.cutpoints( "PgR", status = "ER",
tag.healthy = 0, methods = "Youden", data = datos,
pop.prev = NULL, ci.fit = TRUE, conf.level = 0.95)
```

veamos los resultados

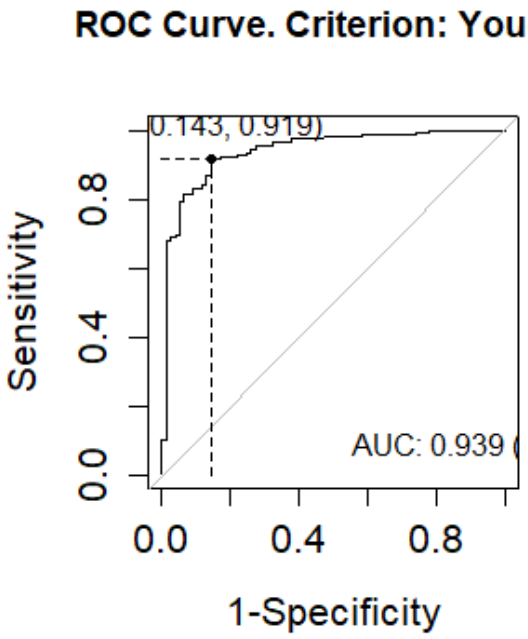
Area under the ROC curve (AUC): 0.939 (0.908, 0.971)

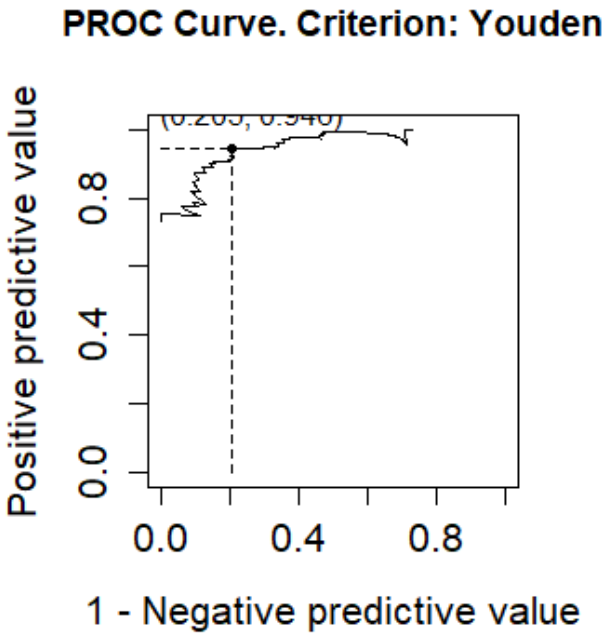
Criterion: Youden
Number of optimal cutoffs: 1

	Estimate	95% CI lower limit	95% CI upper limit
cutoff	1342.900	-	-
S	0.918603	0.8729	0.9515
E	0.8571429	0.7587	0.9264
VPP	0.9458128	0.9014	0.9683
VPN	0.7951807	0.7025	0.8907
LRP	6.430622	3.7153	11.1303
LRN	0.094896	0.0596	0.1510
FP	11.00000	-	-
FN	17.00000	-	-
Optimal criterion	0.7758031	-	-

La variable PgR tiene buena exactitud diagnostica presenta el área bajo la curva ROC (AUC) de 0.939 (IC 95% (0.908.0.971)), el punto de corte optimo se encuentra en 1342.9 basado en el índice Youden (J=0.775), dicho punto de corte presenta una sensibilidad de 91.8% (IC95% (87.2, 95.1)), especificidad de 85.7% (IC95% (75.8,92.6)).

El valor para la sensibilidad que se encontro con el valor umbral de la expresión génica PgR 1342 nos permite creer en el uso de la variable como test (para la determinación de enfermedades y las características determinadas por la prevalencia de una enfermedad que afecta a una población) para el diagnóstico rápido de ER en el sentido de que si no se tiene PgR y obtamos por decidir que no se tiene la enfermedad (cáncer) se tiene un VPN de 0.7951 que es ya una probabilidad alta. Ese valor umbral tiene una especificidad superior, un 85.7%, lo que nos lleva a un VPP para una población de un 94% donde los problemas surgen con ese valor umbral (en los FP). Con la sensibilidad y especificidad que acabamos de estimar se concluye que el punto de corte o valor umbral de la expresión génica se puede emplear como indicador de cáncer de mama en futuros estudios,esta como covariable sujeta a error con la sensibilidad y especificidad que acabamos de estimar.





Ahora utilizaremos otro método el cual es SpEqualSe donde el punto de corte es minimizar $(E(c)-S(c))$ y se toma sensibilidad=especificidad, obtenemos los siguientes resultados:

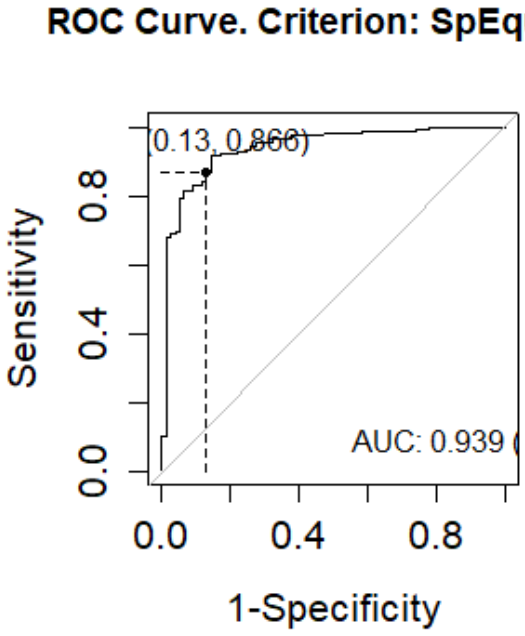
Area under the ROC curve (AUC): 0.939

Criterion: SpEqualSe
Number of optimal cutoffs: 1

	Estimate	95% CI lower limit	95% CI upper limit
cutoff	2414.400	-	-
S	0.8660287	0.8122	0.9090
E	0.8701299	0.7741	0.9359
VPP	0.9476440	0.9025	0.9655
VPN	0.7052632	0.6155	0.8391
LRP	6.668421	3.7313	11.9172
LRN	0.1539670	0.1079	0.2196
FP	10.00000	-	-
FN	28.00000	-	-
Optimal criterion	0.0041011	-	-

La variable PgR tiene buena exactitud diagnostica presenta el área bajo la curva ROC (AUC) de 0.939 (IC 95% (0.908.0.971)). el punto de corte optimo se encuentra 2414.4 basado en SpEqualSe (S=0.0041), dicho punto de corte presenta una sensibilidad de 86.6% (IC95% (81.2, 90.9)), especificidad de 87% (IC95% (77.4,93.5)).

En este punto de corte se toma por igual la sensibilidad y la especificidad.

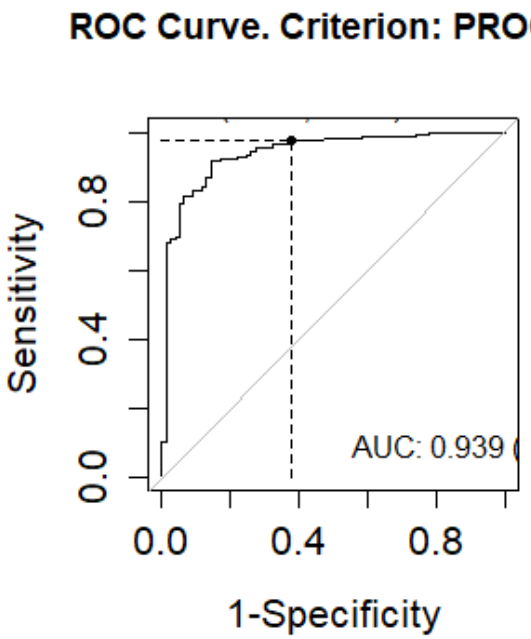


Ahora utilizaremos otro método el cual es PROC01 donde el punto de corte es minimizar $((FFP(c)-1)^2+(FVN(c)-1)^2)$ y se toma la distancia entre la gráfica ROC y el punto (0,1), obtenemos los siguientes resultados:

Area under the ROC curve (AUC): 0.939

Criterion: PROC01
Number of optimal cutoffs: 1

	Estimate	95% CI lower limit	95% CI upper limit
cutoff	472.200	-	-
S	0.97607656	0.9450	0.9921
E	0.62337662	0.5056	0.7313
VPP	0.87553648	0.8129	0.9563
VPN	0.90566038	0.8018	0.9404
LRP	2.59165154	1.9428	3.4571
LRN	0.03837719	0.0158	0.0928
FP	29.00000	-	-
FN	5.00000	-	-
Optimal criterion	0.02439113	-	-



Capítulo 7

Conclusión

En este trabajo es necesario tener en cuenta tres conceptos fundamentales los cuales son; la sensibilidad, especificidad y AUC que son estimadores muestrales de parámetros poblacionales; se dio un marco teórico para la construcción de la curva ROC, teniendo en cuenta los diferentes métodos, a saber: paramétricos, no paramétricos, cada uno de estos métodos se menciona en este trabajo con más profundidad. Se vieron varias formas de cuantificar la capacidad discriminante de un clasificador, se hablo acerca del índice de Youden que nos permite encontrar el punto de corte de mayor sensibilidad y especificidad conjunta, la tasa de verosimilitud nos ha proporcionado una medida de exactitud de clasificación de los individuos con presencia del evento y otra para los individuos con ausencia, el índice de discriminación mide el solapamiento de las funciones de densidades de sanos y enfermos, es decir, es una medida de cuánto le cuesta a la variable de decisión diferenciarlos. Sin embargo, la medida más usada por autores de diferentes campos laborales es el área bajo la curva, esta no sólo nos aporta una medida de la bondad del clasificador, sino que nos permite comparar pruebas. Al respecto, un análisis de curva ROC estadísticamente perfecto carece de todo sentido si los datos utilizados para construir la curva ROC provienen de un estudio metodológicamente deficiente (se menciona en el sesgo de oro).

Con este trabajo se presenta al índice de Youden como un estadístico para ver la exactitud de una prueba diagnóstica que tenga la característica de no depender de la prevalencia, su utilidad dependerá de la medida en que un investigador puede emplearla. Además se muestran distintos enfoques (paramétricos, no paramétricos, semiparamétricos), con los que se construyen intervalos de confianza para el índice de Youden, observando al compararlos entre ellos hay ventajas y desventajas, es posible implementarlos para el análisis de datos en investigación biomédica.

Para la elección del punto de corte podemos hacer uso del índice de Youden, sin embargo, este presenta el problema de maximizar la sensibilidad y especificidad conjuntamente. Una alternativa a este problema es escoger el punto de corte el cual corresponde al par (1-E, S) más cercano al vértice (0,1). En referencia a la obtención del punto de corte optimo por el índice de Youden en una curva ROC, con este trabajo se encontraron diferentes enfoques para encontrar el valor de la prueba donde se hace

máxima diferencia entre la distribución de los valores para población sana y enferma. Los enfoques de estimación de estas distribuciones pueden ser paramétricas y no paramétricas y se han reportado trabajos sobre las distribuciones paramétricas como la Normal, sin embargo, la construcción de distribuciones empíricas también ha sido consideradas.

Las anteriores definiciones se basan en los conceptos de sensibilidad y especificidad, estas medidas poblacionales, pueden estimarse mediante la fracción de verdaderos positivos y la fracción de verdaderos negativos respectivamente, ambas cantidades muestrales. Por tanto, el análisis ROC será más preciso cuanto mejor represente la muestra seleccionada a la población. También hemos construido las curvas ROC correspondientes a los datos de biomarcadores medidos a individuos con y sin cáncer de mama. El sistema usado ha sido R utilizando el paquete pROC permite suavizar la curva mediante diversas distribuciones además de calcular intervalos de confianza para la sensibilidad y especificidad.

En cuanto a la implementación del índice de Youden en los problemas prácticos propuestos, en el paquete de R denominado OptimalcutPoints se pretendió determinar la presencia de cáncer usando como indicador la variable (PgR). Tal punto de corte resultó muy sensible y poco específico, dejando como resultado que la variable PgR puede ser usado para el diagnóstico rápido de estas patologías también se compararon distintos métodos y se observó que el mejor entre ellos para esta base de datos de cáncer de mama fue el índice de Youden y el punto de corte más optimo fue de la misma forma con ese clasificador dejando la puerta abierta a ser usado en estudios posteriores, sujeto a los errores de clasificación.

Fuera del campo de las Matemáticas, las curvas ROC son usadas en diversas áreas, en especial en medicina, cuando ante una enfermedad intentan encontrar el marcador o la medida que la detecte sin necesidad de una prueba de oro (Golden test), es decir, una prueba perfecta sin error en sus resultados, un ejemplo es lo que está pasando actualmente ya que se puede llevar a cabo mediante una curva ROC para clasificar a las personas sanas y enfermas con diferentes variables categóricas tomando en cuenta los distintos errores que pueda tener, esta metodología se puede usar para estudios futuros y así tener una mejor clasificación de las personas.
Otro campo de aplicación es en la meteorología ha favorecido la utilización de estas herramientas en la predicción del tiempo, estimando la probabilidad de ocurrencia en una determinada zona de tornados, huracanes, véase (Stephenson D. (2000)).
Las técnicas ROC han sido recomendadas en el área de Psiquiatría, para evaluar el diagnóstico clínico y la predicción a través de varias variables predictoras por ejemplo la presencia de maltrato infantil (Camasso M. (1995)).

Finalmente, las curvas ROC suponen una herramienta eficiente para medir la bondad de un clasificador y, en caso de ser suficiente, elegir el valor umbral que mejor se ajuste al evento a detectar. Es decir, una prueba más sensible que especifica (preventiva), una más específica que sensible (tamizaje) o una que minimice el o los

resultados erróneos positivos y negativos. Puede realizarse para todo tipo de dato (continuo, discreto, categórico), la sustenta una base teórica que cubre métodos de estimación. Otro punto importante como siguiente paso de este trabajo es desarrollar nuevas metodologías de estimación ya que son necesarias como futuras investigaciones para el avance de la estadística en la medicina.

Bibliografía

[1] Bean,P. (2000): *How good is my test? Supporting decision making using medical test. <http://www.iscpubs.com/articles/ab1/b0012.bea.pdf>.*

[2] Burgueño M.J., García-Bastos J.L. y González-Buitrago J.M. (1993): *Las curvas ROC en la evaluación de las pruebas diagnósticas.* Med. Clin. (Barcelona), 104, págs. 661-670.

[3] Camasso M. J. y Jagannathan R. (1995): Prediction accuracy of the Washington and Illinois risk assessment instruments: an application of receiver operating characteristic curve analysis. Soc. Work Res. 19.

[4] Caro Rubio A. (2017): Adaptación del software R para estudios de fiabilidad. (Grado en ingeniería mecánica), págs. 1-54.

[5] Estrada Álvarez J. M. (2015): *El índice de Youden y su aplicación a la determinación del punto de corte en un test cuantitativo.* (Máster en Estadística Aplicada), Universidad de Granada, págs. 9-10.

[6] Faraggi,D. y Reiser, B. (2002): *Estimation of the area under the ROC curve.* Statistics in Medicine. 21, págs. 3093-3096.

[7] Fawcett, T. (2006): *An introduction to ROC analysis.* 27, págs. 3861-874.

[8] Franco Nicolás Manuel Vivo Molina Juana María (2007): Análisis de curvas Roc principios básicos y aplicaciones. La muralla. Madrid. págs. 17-87.

[9] Green D.M. y Swets J.A. (1996) *Signal Detection Theory and Psychophysics.* John Wiley and Sons, New York.

[10] Jokiel-Rokita A. y Pulit M. (2013): *Nonparametric estimation of the ROC curve based on smoothed empirical distribution functions, Statistics and Computing*, **23**, págs. 703-712.

[11] Lloyd Chris J. (2001): Estimation of a convex ROC curve. Statist. Probab. 44, págs. 100-110.

[12] López Raton M., Rodríguez Álvarez M. X., Cadarso Suarez C. y Gude Sampedro F. (2014): *Optimal Cutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software.*

[13] Luciani D., Facchini R. y Bertonoli G. (2000): The (ROC) curve , págs. 1-20.

[14] Lusted L.B. (1971): *Decision-making studies in patient management.* New Engl. J. Med. 284.

[15] Morillo Paulina A. (2016): Estudio y Desarrollo de una Librería en R para Evaluar las Prestaciones de un Clasificador, págs. 1-38.

[16] Ochoa Sangrador C. y Orejas G. (1960): *Epidemiología y metodología científica aplicada a la pediatría (IV): Pruebas diagnósticas*, **8**, págs. 301-314.

[17] Passas Martínez M. (2012): *Cálculo del umbral (GRACE score) en el síndrome coronario agudo mediante curva ROC.* (Trabajo Fin de Máster.) Universidad de Granada, págs. 1-58.

[18] Pepe Margaret S. y Patrick J. (2000): *Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker.* New York, págs. 237-344.

[19] Pulit M. (2015): A new method of kernel-smoothing estimation of the ROC curve, págs. 603-634.

[20] Quintela del Río A. y Estévez Pérez G. (2012): Nonparametric Kernel Distribution Function Estimation with kerdiest: An R Package for Bandwidth Choice and Applications. Journal of Statistical Software. 50, págs. 1-21.

[21] Ron E., Domrachev M. y Lash A.E. (2002): *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* In: Nucleic Acids Res. 30, págs. 207-210.

[22] Robin X. (2011): *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* Swiss Institute of Bioinformatics, Geneva, Switzerland.

[23] Stephenson D. B. (2000): Use of the "*Odds Ratio*" for Diagnosing Forecast Skill. Weather Forecast. 15, págs. 221-232.

[24] Torres Ortiz A. (2010): ROC para Datos de Supervivencia. Aplicación a Datos Biomédicos. (Máster en Técnicas Estadísticas), Universidad de Santiago de Compostela, págs. 1-71.

[25] Wang Y., Klijn J.G. y Zhang Y. (2005): *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* In: Lance. 365, págs. 671-679.

[26] Woo S. y Henderson D. (2015): *Dichotomization of Continuous Biomarkers.* Axio Research. Seattle, págs. 1-8.

[27] Zhou X.H., Obuchowski N.A. y McClish D.K. (2002): *Statistical Methods in Diagnostic Medicine.* Wiley Inter-Science, New York.

[28] Zou K.H., Tempany C.M., Fielding J.R. y Silverman S.G. (1998): *Original smooth reciever operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones.* Acad. Radiol. 5, págs. 680-687.