Técnicas diagnósticas

(ver 2019-2020)

Elisa M. Molanes-López y Emilio Letón

20-Oct-2019

Página legal

ÍNDICES DE RIESGO

Madrid, versión 2019-2020

ISBN electrónico: xxx.
Edición digital (epub): xxx.

Prefacio

Este material está diseñado con el paradigma del grupo de innovación docente miniXmodular de generar material en formato mini y modular, según se puede ver en la página web www.minixmodular.ia.uned.es, donde se introduce, entre otros, el concepto de mini-libro electrónico modular.

En este mini-libro se contemplan dos variables X e Y donde X es la variable explicativa e Y es la variable respuesta. En este contexto de técnicas diagnósticas, la variable X se denomina variable de diagnóstico o prueba de diagnóstico ya que su objetivo es ayudar a diagnosticar la presencia o ausencia de una enfermedad recogida en la variable Y, la cual está codificada como 1 ("Sí enfermo") y 0 ("No enfermo" o "Sano"). En el mini-capítulo 1 se tratará el caso de que la variable diagnóstico X sea dicotómica y en el mini-capítulo 2 el caso de que X sea continua.

Se recomienda refrescar los conceptos básicos de intervalos de confianza (IC) y contraste de hipótesis (CH) en, por ejemplo, [7].

Por último, conviene mencionar que a lo largo de este mini-libro se presenta código en el lenguaje R [6] para realizar los cálculos estadísticos. Existen numerosas páginas web y libros que introducen este lenguaje; dos referencias clásicas que combinan Estadística y R son [1] y [2]. No obstante, conviene señalar que el código que se utiliza es muy sencillo y autoexplicativo.

Mini-capítulo 1

Diagnóstico dicotómico

En este mini-capítulo se estudia el caso de diagnóstico dicotómico, con lo que se está en el contexto de X variable diagnóstico dicotómica e Y variable enfermedad dicotómica, por lo que el esquema es $D \leftarrow D$. En este contexto, $X = 1$ indica que la prueba diagnóstica afirma que el individuo está enfermo y $X = 0$ indica que la prueba diagnóstica dice que el individuo no está enfermo, mientras que $Y = 1$ indica que el individuo está realmente enfermo e $Y = 0$ que está realmente sano. Conviene tener presente que en el caso de índices de riesgo con esquema $D \leftarrow D$, se tiene que $X = 1$ indica que el individuo está expuesto a un factor de riesgo y que $X = 0$ indica que el individuo no está expuesto a un factor de riesgo.

Al tomar una muestra en el contexto de diagnóstico dicotómico, los datos experimentales tendrán el patrón dado en la tabla 1.1, donde el número de individuos con perfil $Y = 1$ y $X = 1$ es a, con perfil $Y = 1$ y $X = 0$ es b, con perfil $Y = 0$ y $X = 1$ es c y con perfil $Y = 0$ y $X = 0$ es d.

Y	X
1	1
...	...
1	1
1	0
...	...
1	0
0	1
...	...
0	1
0	0
...	...
0	0

Tabla 1.1: Datos genéricos de diagnóstico del esquema $D \leftarrow D$

A partir de los datos experimentales se construye la tabla de datos cruzados genérica dada en la tabla 1.2, donde los marginales son $r_1 = a + b$, $r_0 = c + d$, $s_1 = a + c$, $s_0 = b + d$ y $n = r_1 + r_0 = s_1 + s_0$ es el tamaño muestral, con independencia del diseño (si se fijan los marginales r_1 y r_0, los marginales s_1 y s_0 o el total n). En la tabla 1.2, se indica con a a los individuos clasificados correctamente al afirmar (de forma positiva) con la ayuda de la prueba diagnóstica X que el individuo es "Sí enfermo" (VP = verdaderos positivos), con b los clasificados incorrectamente al afirmar (de forma negativa) con la ayuda de X que el individuo es "No enfermo" (FN = falsos negativos), con c los clasificados incorrectamente al afirmar (positivamente) con la ayuda de X que el individuo es "Sí enfermo" (FP = falsos positivos) y con d los individuos clasificados correctamente al afirmar (negativamente) con la ayuda de X que el individuo es "No enfermo" (VN = verdaderos negativos).

	$X = 1$	$X = 0$	
$Y = 1$	$a = VP$	$b = FN$	r_1
$Y = 0$	$c = FP$	$d = VN$	r_0
	s_1	s_0	n

Tabla 1.2: Tabla con datos genéricos de diagnóstico del esquema $D \leftarrow D$

En este mini-capítulo se estudia, asociados a los datos de la tabla 1.2, los parámetros poblacionales sensibili-

dad y especificidad, valores predictivos y razones de verosimilitudes dicotómicas ("likelihood ratio" dicotómicos).

1.1. Sensibilidad y especificidad

El parámetro poblacional θ sensibilidad poblacional está dado por

$$\theta = se = \pi'_{11} = P(X = 1|Y = 1)$$

que representa "el acierto en enfermos" y el parámetro poblacional θ especificidad poblacional está dado por

$$\theta = es = \pi'_{00} = P(X = 0|Y = 0)$$

que representa "el acierto en sanos". También se suelen definir las denominadas probabilidades de clasificación dadas por la fracción de verdaderos positivos (*fvp*) y la fracción de falsos positivos (*ffp*) como

$$\begin{aligned} fvp &= P(X = 1|Y = 1) = se \\ ffp &= P(X = 1|Y = 0) = 1 - es, \end{aligned}$$

según se puede ver en [5].

Al ser *se* y *es* proporciones, estarán comprendidas entre 0 y 1, y si se expresan en tanto por ciento entre 0 % y 100 %. El hecho de que sean proporciones facilita la construcción de IC y CH para dichos parámetros. Para estimar el parámetro *se* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi}'_{11}$ (es decir, $\widehat{\Theta} = \widehat{SE} = \widehat{\Pi}'_{11}$), siendo $\widehat{\Pi}'_{11}$ la proporción muestral entendida como v.a. ya que variará de muestra a muestra y que aplicado a la muestra de estudio proporciona $\widehat{se} = \frac{a}{r_1}$. Por otra parte, para estimar el parámetro *es* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi}'_{00}$ (es decir, $\widehat{\Theta} = \widehat{ES} = \widehat{\Pi}'_{00}$) y que aplicado a la muestra de estudio proporciona $\widehat{es} = \frac{d}{r_0}$.

La sensibilidad y la especificidad se pueden interpretar también en un contexto de CH, en el sentido de que en un CH se puede considerar la variable X como "decisión que se toma en el CH", donde $X = 1$ representa concluir la hipótesis alternativa H_1 y $X = 0$ no concluir H_1. Por otra parte, la variable Y sería el "estado real de la naturaleza", donde $Y = 1$ representa que la hipótesis alternativa H_1 es realmente cierta e $Y = 0$ que es la hipótesis H_0 la que realmente es cierta. En este contexto de CH, el error de tipo I sería decir que hay diferencias cuando en realidad no las hay, es decir $\alpha = P(\text{concluir } H_1|\text{en realidad } H_0 \text{ es cierta}) = 1 - es$ y el error de tipo II sería decir que no hay diferencias cuando en realidad las hay, es decir $\beta = P(\text{concluir } H_0|\text{en realidad } H_1 \text{ es cierta}) = 1 - se$, con lo que la potencia sería $1 - \beta = se$.

Se puede comprobar que *se* y *es* no dependen de la prevalencia (la probabilidad de enfermedad, $\pi = P(Y = 1)$), por lo que son útiles para reflejar la bondad de una prueba diagnóstica.

Se observa que, aunque siempre $P(X = 1|Y = 1) + P(X = 0|Y = 1) = 1$, sin embargo en general $P(X = 1|Y = 1) + P(X = 0|Y = 0) = se + es \neq 1$. Asociadas a la sensibilidad y especificidad, se definen el índice de Youden publicado en 1950 y la eficacia a través de

$$\text{Youden} = se + es - 1$$

$$\text{eficacia} = \pi se + (1 - \pi)es$$

Ejemplo

1. La probabilidad de cometer una predicción errónea es

$$P(error) = \pi(1 - fvp) + (1 - \pi)ffp.$$

SOLUCIÓN:

Se tiene que

$$\begin{aligned} P(error) &= P(X \neq Y) = P((X = 0 \cap Y = 1) \cup (X = 1 \cap Y = 0)) \\ &= P(X = 0 \cap Y = 1) + P(X = 1 \cap Y = 0) \\ &= P(Y = 1)P(X = 0|Y = 1) + P(Y = 0)P(X = 1|Y = 0) \\ &= P(Y = 1)(1 - P(X = 1|Y = 1)) + (1 - P(Y = 1))P(X = 1|Y = 0) \\ &= \pi(1 - fvp) + (1 - \pi)ffp \end{aligned}$$

2. Sea X la variable diagnóstico e Y la variable enfermedad con los siguientes datos experimentales dados en la tabla 1.3. Se pide calcular:

	$X=1$	$X=0$	
$Y=1$	10	30	40
$Y=0$	12	93	105
	22	123	145

Tabla 1.3: Tabla con datos observados de diagnóstico del esquema $D \leftarrow D$

- El $\widehat{or}$, su IC y su significación estadística.
- A la vista del apartado anterior, ¿qué se puede decir sobre la bondad diagnóstica de X?

SOLUCIÓN:

- La siguiente sintaxis en R proporciona $\widehat{or} = 2.5833$ con $IC95\,\%(or) = (1.0145, 6.5782)$ y su p-valor=$0.0418 < 0.05$ (resultado significativo).

```
rm(list=ls())

datos=read.table('d_d_4.txt',header=T)
attach(datos)

ind1=which(exp==1);
ind0=which(exp==0);
ind11=which(rta==1 & exp==1);
ind10=which(rta==1 & exp==0);
a=length(rta[ind11]); a
b=length(rta[ind10]); b
s1=length(rta[ind1]); s1
s0=length(rta[ind0]); s0
c=s1-a; c
d=s0-b; d
r1=a+b; r1
r0=c+d; r0
n=a+b+c+d; n

alfa=0.05
or=(a*d)/(b*c); or
ee.lnor=sqrt(1/a+1/b+1/c+1/d)
ic1=exp(log(or)-qnorm(1-alfa/2)*ee.lnor); ic1
ic2=exp(log(or)+qnorm(1-alfa/2)*ee.lnor); ic2

tabla=table(-rta,-exp); tabla
chisq.test(tabla,correct=FALSE)
```

- En un contexto de índices de riesgo, se tendría que al ser $\widehat{or} = 2.5833 > 1$ y significativo (p-valor= $0.0418 < 0.05$), la prueba de diagnóstico estaría funcionando como "factor de riesgo" (presencia de prueba diagnóstico positiva asociada a presencia de enfermedad). Sin embargo esta prueba diagnóstica es muy anómala como se puede observar comparando la diagonal principal de la tabla con los valores fuera de ella, habiendo más individuos con $X = 1$ y sanos que individuos con $X = 1$ y enfermos. Esta situación no es la que se espera para una buena prueba diagnóstica. Este ejemplo proporciona de forma intuitiva la razón por la que hay que considerar nuevos parámetros que sean válidos en las técnicas diagnósticas como la sensibilidad y la especificidad.

1.2. Valor predictivo positivo y valor predictivo negativo

El parámetro poblacional θ valor predictivo positivo poblacional *vpp* está dado por

$$\theta = vpp = \pi_{11} = P(Y = 1|X = 1)$$

que representa "el acierto en los individuos positivos según la prueba diagnóstica" y el parámetro poblacional θ valor predictivo negativo poblacional *vpn* está dado por

$$\theta = vpn = \pi_{00} = P(Y = 0|X = 0)$$

que representa "el acierto en los individuos negativos según la prueba diagnóstica".

Al ser *vpp* y *vpn* proporciones, estarán comprendidas entre 0 y 1, y si se expresan en tanto por ciento entre 0 % y 100 %. El hecho de que sean proporciones facilita la construcción de IC y CH para dichos parámetros. Estas probabilidades *vpp* y *vpn* también reciben el nombre de probabilidades de predicción (ver, por ejemplo, [5]). Para estimar el parámetro *vpp* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi}_{11}$ (es decir, $\widehat{\Theta} = \widehat{VPP} = \widehat{\Pi}_{11}$), siendo $\widehat{\Pi}_{11}$ la proporción muestral entendida como v.a. ya que variará de muestra a muestra y que para la muestra de estudio proporciona $\widehat{vpp} = \frac{a}{s_1}$. Por otra parte, para estimar el parámetro *vpn* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi}_{00}$ (es decir, $\widehat{\Theta} = \widehat{VPN} = \widehat{\Pi_{00}}$) que, aplicado a la muestra de estudio, proporciona $\widehat{vpn} = \frac{d}{s_0}$.

La utilidad de *vpp* y *vpn* es práctica ya que son probabilidades de predicción. Se puede comprobar que *vpp* y *vpn* dependen de la prevalencia, por lo que no son útiles para reflejar la bondad de una prueba diagnóstica a diferencia de la *se* y la *es*.

Ejemplo

1. En el contexto de X variable diagnóstico dicotómica e Y variable enfermedad dicotómica, el odds a priori de enfermedad es $\frac{\pi}{1-\pi}$ y el odds a posteriori de enfermedad es $\frac{vpp}{1-vpp}$ en el caso de $X = 1$ y $\frac{1-vpn}{vpn}$ en el caso de $X = 0$.

   SOLUCIÓN:

   Al ser $\pi = P(Y = 1)$ la probabilidad de enfermedad, el odds a priori de enfermedad es $\frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi}{1-\pi}$. Por otra parte el odds a posteriori de enfermedad es $\frac{P(Y=1|X)}{1-P(Y=1|X)}$, que para $X = 1$ es $\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \frac{vpp}{1-vpp}$ y para $X = 0$ es $\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \frac{1-vpn}{vpn}$.

1.3. Razón de verosimilitud positiva y razón de verosimilitud negativa

El parámetro poblacional θ razón de verosimilitud positiva poblacional *lrp* y el parámetro poblacional θ razón de verosimilitud negativa poblacional *lrn* están dados por

$$\theta = lrp = \frac{\pi'_{11}}{\pi'_{10}} = \frac{P(X = 1|Y = 1)}{P(X = 1|Y = 0)} = \frac{se}{1 - es}$$

$$\theta = lrn = \frac{\pi'_{01}}{\pi'_{00}} = \frac{P(X = 0|Y = 1)}{P(X = 0|Y = 0)} = \frac{1 - se}{es}.$$

Según se puede ver en, por ejemplo, [8], *lrp* representa "la ganancia de información cuando la prueba diagnóstica da positivo" y *lrn* representa "la ganancia de información cuando la prueba diagnóstica da negativo". Para estimar el parámetro *lrp* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\frac{\widehat{\Pi}'_{11}}{\widehat{\Pi}'_{10}}$ (es decir, $\widehat{\Theta} = \widehat{LRP} = \frac{\widehat{\Pi}'_{11}}{\widehat{\Pi}'_{10}}$), siendo $\widehat{\Pi}'_{11}$ y $\widehat{\Pi}'_{10}$ las proporciones muestrales entendidas como v.a. ya que variarán de muestra a muestra, y que para la muestra de estudio proporciona $\widehat{lrp} = \frac{\widehat{se}}{1-\widehat{es}}$. Por otra parte, para estimar el parámetro *lrn* se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\frac{\widehat{\Pi}'_{01}}{\widehat{\Pi}'_{00}}$ (es decir, $\widehat{\Theta} = \widehat{LRN} = \frac{\widehat{\Pi}'_{01}}{\widehat{\Pi}'_{00}}$), que, aplicado a la muestra de estudio proporciona $\widehat{lrn} = \frac{1-\widehat{se}}{\widehat{es}}$.

Estas razones de verosimilitud son realmente un cociente de verosimilitudes, que en el caso discreto, son probabilidades

$$lr(x) = \frac{P_1(X = x)}{P_0(X = x)},$$

donde $P_1(X = x) = P(X = x|Y = 1)$ y $P_0(X = x) = P(X = x|Y = 0)$.

Conviene señalar que las razones de verosimilitud no son proporciones sino razones (un cociente de proporciones en este caso) y que están comprendidos en el intervalo $(0, \infty)$. Además, se observa directamente de su definición (en términos de *se* y *es*) que no dependen de la prevalencia (ya que ni *se* y *es* dependen de ella).

Estas razones de verosimilitud se interpretan de la forma siguiente:

- Si $lrp > 1$, es más probable que un resultado positivo en la prueba diagnóstica se dé en un sujeto enfermo que en uno sano.
- Si $lrn < 1$, es menos probable que un resultado negativo en la prueba diagnóstica se dé en un sujeto enfermo que en uno sano.

Al ser $\widehat{LRP}$ y $\widehat{LRN}$ v.a. asimétricas (su rango es $(0,\infty)$), su distribución es no normal. Es conveniente, por tanto, considerar una transformación que consiga normalidad con media y varianza teóricas que se puedan estimar fácilmente. Se puede demostrar que la transformación logaritmo neperiano (Ln) consigue este propósito. A efectos teóricos, el parámetro θ de interés es $Ln(lrp)$ y $Ln(lrn)$ y su estimador $\widehat{\Theta}$ es $Ln(\widehat{LRP})$ y $Ln(\widehat{LRN})$, respectivamente.

En este contexto, y aplicando el teorema de Taylor para v.a. (ver, por ejemplo, [3] y [4]), se pueden construir intervalos de confianza al $(1-\alpha)\,\%$ para $Ln(lrp)$ y $Ln(lrn)$ dados por

$$\begin{aligned} IC(1-\alpha)\,\%(Ln(lrp)) &= \Big(\widehat{E}[Ln(\widehat{LRP})] \mp z_{1-\alpha/2}\widehat{EE}[Ln(\widehat{LRP})]\Big), \\ IC(1-\alpha)\,\%(Ln(lrn)) &= \Big(\widehat{E}[Ln(\widehat{LRN})] \mp z_{1-\alpha/2}\widehat{EE}[Ln(\widehat{LRN})]\Big), \end{aligned}$$

con

$$\widehat{E}[Ln(\widehat{LRP})] = Ln(lrp) = Ln\Big(\frac{\widehat{se}}{1-\widehat{es}}\Big),$$

$$\widehat{EE}[Ln(\widehat{LRP})] = \sqrt{\frac{1}{r_1}\frac{1-\widehat{se}}{\widehat{se}} + \frac{1}{r_0}\frac{\widehat{es}}{1-\widehat{es}}},$$

$$\widehat{E}[Ln(\widehat{LRN})] = Ln(lrn) = Ln\Big(\frac{1-\widehat{se}}{\widehat{es}}\Big),$$

$$\widehat{EE}[Ln(\widehat{LRN})] = \sqrt{\frac{1}{r_1}\frac{\widehat{se}}{1-\widehat{se}} + \frac{1}{r_0}\frac{1-\widehat{es}}{\widehat{es}}}.$$

Deshaciendo la transformación del logaritmo neperiano se tiene que

$$\begin{aligned} IC(1-\alpha)\,\%(lrp) &= \Big(exp\Big(Ln\Big(\frac{\widehat{se}}{1-\widehat{es}}\Big) \mp z_{1-\alpha/2}\sqrt{\frac{1}{r_1}\frac{1-\widehat{se}}{\widehat{se}} + \frac{1}{r_0}\frac{\widehat{es}}{1-\widehat{es}}}\Big)\Big), \\ IC(1-\alpha)\,\%(lrn) &= \Big(exp\Big(Ln\Big(\frac{1-\widehat{se}}{\widehat{es}}\Big) \mp z_{1-\alpha/2}\sqrt{\frac{1}{r_1}\frac{\widehat{se}}{1-\widehat{se}} + \frac{1}{r_0}\frac{1-\widehat{es}}{\widehat{es}}}\Big)\Big). \end{aligned}$$

Ejemplo

1. En el contexto de X variable diagnóstico dicotómica e Y variable enfermedad dicotómica, los odds a posteriori de enfermedad están dados por $lrp \cdot$ odds a priori de enfermedad en el caso de que $X=1$ y por $lrn \cdot$ odds a priori de enfermedad en el caso de que $X=0$.

   SOLUCIÓN:

   En primer lugar, se tiene que el odds a posteriori de enfermedad de resultado $X=1$ es

   $$\begin{aligned} \frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} &= \frac{\frac{P(X=1|Y=1)P(Y=1)}{P(X=1)}}{\frac{P(X=1|Y=0)P(Y=0)}{P(X=1)}} = \frac{P(X=1|Y=1)}{P(X=1|Y=0)}\frac{P(Y=1)}{P(Y=0)} \\ &= lrp \cdot \text{odds a priori de enfermedad} \end{aligned}$$

   En segundo lugar, se tiene que el odds a posteriori de enfermedad de resultado $X=0$ es

   $$\begin{aligned} \frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} &= \frac{\frac{P(X=0|Y=1)P(Y=1)}{P(X=0)}}{\frac{P(X=0|Y=0)P(Y=0)}{P(X=0)}} = \frac{P(X=0|Y=1)}{P(X=0|Y=0)}\frac{P(Y=1)}{P(Y=0)} \\ &= lrn \cdot \text{odds a priori de enfermedad} \end{aligned}$$

   Con la ayuda de las fórmulas anteriores se tiene que lrp se puede interpretar como la ganancia de información cuando X es positivo ($X=1$) y lrn como la ganancia de información cuando X es negativo ($X=0$).

Mini-capítulo 2

Diagnóstico continuo

En este mini-capítulo se estudia el caso de diagnóstico continuo, con lo que se está en el contexto de X variable diagnóstico continua (que recibe el nombre de biomarcador) e Y variable enfermedad dicotómica, por lo que el esquema es $D \leftarrow C$. Se suele asumir, sin pérdida de generalidad (ver, por ejemplo, [5]), que valores altos de X están asociados con la presencia de enfermedad ($Y = 1$) y que valores bajos de X corresponden a la ausencia de enfermedad ($Y = 0$). Un ejemplo de esta situación sería estudiar si el biomarcador dado por el PSA (X) sirve para diagnosticar el cáncer de próstata (Y).

Una forma de estudiar esta situación es dicotomizar la variable X, a través de una nueva X^*, para así estar en el supuesto del mini-capítulo anterior. La variable X^* se define como

$$X^* = \begin{cases} 1 & \text{si} \quad X \geq c \\ 0 & \text{si} \quad X < c \end{cases}$$

donde c es un punto de corte o umbral que habrá que determinar de forma óptima. Utilizando esta X^*, un individuo se clasificará como positivo o "enfermo" ($X^* = 1$) si el biomarcador X supera ese punto de corte y como negativo o "sano" en caso contrario ($X^* = 0$). Por lo tanto, para cada c es posible definir la sensibilidad $se(c)$, la fracción de verdaderos positivos $fvp(c)$, la especificidad $es(c)$ y la fracción de falsos positivos $ffp(c)$ mediante

$$\begin{aligned} se(c) &= P(X^* = 1|Y = 1) = P(X \geq c|Y = 1), \\ fvp(c) &= se(c), \\ es(c) &= P(X^* = 0|Y = 0) = P(X < c|Y = 0), \\ ffp(c) &= P(X^* = 1|Y = 0) = 1 - es(c). \end{aligned}$$

A partir de los valores $ffp(c)$ y $fvp(c)$ se define la curva roc ("receiver operating characteristic") que se estudiará en la mini-sección 2.1.

2.1. Curvas *roc*

La curva roc tiene su origen en la Teoría de la señal en los años 1950-1960, en el contexto de los radares con el objetivo de distinguir entre señal y ruido (en concreto, entre la llegada o no de misiles). Más adelante se introdujo en radiología con Hanley y McNeil en los años 1982 y 1983 y hoy en día es muy utilizada en Investigación clínica y también en el estudio de la bondad del comportamiento de modelos estadísticos.

La curva roc poblacional se define como

$$roc(\cdot) = \{(ffp(c), fvp(c)), c \in (-\infty, +\infty)\},$$

es decir, representando para distintos puntos de corte, en el eje de abscisas la $ffp(c)$ y en el eje de ordenadas la $fvp(c)$. Por tanto,

$$roc(\cdot) = \{(t, roc(t)), t \in (0, 1)\},$$

siendo $roc(t) = fvp(c)$ con c tal que $ffp(c) = t$. A medida que c crece, $fvp(c)$ y $ffp(c)$ decrecen y a medida que c decrece, $fvp(c)$ y $ffp(c)$ crecen.

El área bajo la curva roc poblacional se denota por el parámetro poblacional auc ("area under the curve") y está dada por

$$auc = \int_0^1 roc(t)dt$$

y constituye un índice de resumen de la bondad del biomarcador X. El *auc* se puede estimar por el método trapezoidal de suma de áreas de trapecios con

$$\widehat{auc} = \sum_{i=1}^{k} (\mathit{ffp}_i - \mathit{ffp}_{i-1}) \frac{se_i + se_{i-1}}{2}$$

donde k indica el número de puntos de corte considerados. Se demuestra que son equivalentes el *auc* y el test U de Mann-Whitney.

Conviene señalar que un biomarcador con capacidad de discriminación nula tiene un $auc = 0.5$. Esto último se daría, por ejemplo, si se utilizara un procedimiento aleatorio para clasificar a los individuos.

Un aspecto muy importante en el contexto de curvas *roc* es determinar un punto de corte óptimo para conseguir maximizar la capacidad diagnóstica de un biomarcador en la práctica. Es importante observar que no es posible conseguir maximizar $se(c)$ y $es(c)$ a la vez, por lo que hay que elegir una situación de equilibrio o compromiso con el punto de corte c que se utilice. Hay dos métodos básicos para determinar un punto de corte óptimo (ver, por ejemplo, [5]): la esquina noroeste (se toma el c que hace que el punto de la curva *roc* esté más cerca del punto $(0, 1)$) y el índice de Youden (se toma el c que hace máxima la distancia vertical de la curva *roc* a la diagonal).

Ejemplo

1. Si se define S la función de supervivencia de una v.a. como el complementario a uno de la función de distribución y se denotan por S_1 y S_0 las funciones de supervivencia de X para el grupo de enfermos y sanos, respectivamente, se tiene que
$$roc(t) = S_1(S_0^{-1}(t)),$$
con $t \in (0, 1)$. Además, se verifica que $auc = P(X_1 > X_0)$.

   SOLUCIÓN:

   Por una parte $S_1(x) = 1 - F_1(x) = 1 - P(X \leq x|Y = 1) = P(X > x|Y = 1) = P(X_1 > x)$ y $S_0(x) = 1 - F_0(x) = 1 - P(X \leq x|Y = 0) = P(X > x|Y = 0) = P(X_0 > x)$. Por otra parte, de la definición de curva *roc* como $roc(\cdot) = \{(t, roc(t)), t \in (0, 1)\}$, siendo $roc(t) = \mathit{fvp}(c)$ con c tal que $\mathit{ffp}(c) = t$ se sigue que

$$t = \mathit{ffp}(c) = \mathit{ffp}(c) = P(X \geq c|Y = 0) = P(X > c|Y = 0) = S_0(c)$$

   con lo que $c = S_0^{-1}(t)$ y, por lo tanto, $roc(t) = \mathit{fvp}(c) = P(X \geq c|Y = 1) = P(X > c|Y = 1) = S_1(c) = S_1(S_0^{-1}(t))$.

   A partir de la última expresión, se tiene que

$$\begin{aligned} auc &= \int_0^1 roc(t)dt = \int_0^1 S_1(S_0^{-1}(t))dt = \int_{y=-\infty}^{+\infty} S_1(y)dS_0(y) \\ &= \int_{y=-\infty}^{+\infty} P(X_1 > y)f_0(y)dy \overset{indep.}{=} \int_{y=-\infty}^{+\infty} P(\{X_1 > y\} \cap \{X_0 = y\})dy = P(X_1 > X_0). \end{aligned}$$

   Es decir, el *auc* representa la probabilidad de que dados un par de individuos, uno enfermo y otro sano, éstos estén correctamente ordenados según el biomarcador X, resultado probado por Bamber en 1975.

2. Si X_1 es $N(\mu_1, \sigma_1)$ y X_0 es $N(\mu_0, \sigma_0)$, entonces se verifica que

$$\begin{aligned} roc(t) &= \Phi(a + b\Phi^{-1}(t)) \\ auc &= \Phi\Big(\frac{a}{\sqrt{1 + b^2}}\Big) \end{aligned}$$

   donde $a = \dfrac{\mu_1 - \mu_0}{\sigma_1}$, $b = \dfrac{\sigma_0}{\sigma_1}$ y Φ es la función de distribución de la $N(0, 1)$.

   SOLUCIÓN:

   Por una parte

$$\begin{aligned} \mathit{fvp}(c) &= P(X \geq c|Y = 1) = P(X_1 \geq c) = P(X_1 \leq -c) = \Phi\Big(-\frac{c - \mu_1}{\sigma_1}\Big) = \Phi\Big(\frac{\mu_1 - c}{\sigma_1}\Big), \\ \mathit{ffp}(c) &= P(X \geq c|Y = 0) = P(X_0 \geq c) = P(X_0 \leq -c) = \Phi\Big(-\frac{c - \mu_0}{\sigma_0}\Big) = \Phi\Big(\frac{\mu_0 - c}{\sigma_0}\Big). \end{aligned}$$

Por otra parte, dado que $roc(t) = fvp(c)$ con c tal que $ffp(c) = t$, se tiene que

$$\begin{aligned} \Phi\Big(\frac{\mu_0 - c}{\sigma_0}\Big) &= t \Rightarrow \Phi^{-1}(t) = \frac{\mu_0 - c}{\sigma_0} \Rightarrow c = \mu_0 - \sigma_0\Phi^{-1}(t) \\ roc(t) &= fvp(c) = \Phi\Big(\frac{\mu_1 - c}{\sigma_1}\Big) = \Phi\Big(\frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(t)}{\sigma_1}\Big) = \Phi(a + b\Phi^{-1}(t)). \end{aligned}$$

Por último, dado que $auc = P(X_1 > X_0) = P(X_1 - X_0 > 0) = P(W > 0)$, donde $W = X_1 - X_0$ es una $N(\mu_1 - \mu_0, \sqrt{\sigma_1^2 + \sigma_0^2})$, se tiene que

$$\begin{aligned} auc &= 1 - P(W \le 0) = 1 - P\Big(\frac{W - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}} \le \frac{-(\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\Big) \\ &= 1 - P\Big(Z \le \frac{-(\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\Big) = 1 - \Phi\Big(\frac{\mu_0 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_0^2}}\Big) = \Phi\Big(-\frac{\mu_0 - \mu_1}{\sqrt{\sigma_1^2 + \sigma_0^2}}\Big) = \Phi\Big(\frac{a}{\sqrt{1 + b^2}}\Big). \end{aligned}$$

3. Se ha realizado un estudio para ver la bondad diagnóstica de un biomarcador a la hora de detectar una enfermedad. Han participado en el estudio 17 individuos. Los datos experimentales se dan en la tabla 2.1. ¿Cuánto vale la $\widehat{auc}$? ¿Qué relación tiene con la U de Mann-Whitney?

rta	*exp*
1	10
1	10
1	30
1	30
1	30
1	50
1	70
0	10
0	20
0	20
0	30
0	40
0	40
0	40
0	40
0	60
0	70

Tabla 2.1: Datos observados del esquema $D \leftarrow C$

SOLUCIÓN:

Para calcular la $\widehat{auc}$ hay que fijar varios puntos de corte c y determinar para cada punto de corte $ffp(c)$ y $fvp(c)$ según se muestra en la tabla 2.2, donde en la última columna se recogen las áreas trapezoidales de la curva *roc*.

c	$ffp(c)$	$fvp(c)$	$\widehat{auc(c)}$
9	1	1	0.0857
15	9/10	5/7	0.1429
25	7/10	5/7	0.0500
35	6/10	2/7	0.1143
45	2/10	2/7	0.0000
55	2/10	1/7	0.0143
65	1/10	1/7	0.0071
71	0	0	0.0000

Tabla 2.2: Áreas trapezoidales para datos observados del esquema $D \leftarrow C$

Sumando todas las áreas trapezoidales de la última columna de la tabla 2.2, se obtiene que $\widehat{auc} = 0.4143$. Para ver la relación con el test U de Mann-Whitney habría que intercambiar los papeles de variable

explicativa y variable respuesta y se podría comprobar que $\frac{U_{01}}{n_0 n_1} = \frac{29}{10 \cdot 7} = 0.4143 = \widehat{auc}$. Al ser este valor menor que 0.5 se tiene que este biomarcador X es peor que un biomarcador aleatorio.

Por último, se muestra a continuación la sintaxis en R para dibujar la curva $\widehat{roc}$ y calcular el $\widehat{auc}$, suponiendo que los datos están en el fichero d_c_1.txt.

```
rm(list=ls())

datos=read.table('d_c_1.txt',header=T)
attach(datos)

library(pROC)

roc_obj <- roc(datos$rta, datos$exp)
# Imprimir el AUC
auc(roc_obj)
# Dibujar la curva ROC
## Opción 1
plot(roc_obj, print.auc=TRUE)
## Opción 2
roc_df <- data.frame(
  fvp=rev(roc_obj\$sensitivities),
  ffp=rev(1 - roc_obj\$specificities))
plot(0:10/10, 0:10/10, type='n', xlab="ffp", ylab="fvp")
abline(h=0:10/10, col="lightblue")
abline(v=0:10/10, col="lightblue")
abline(coef = c(0,1), col="lightblue")
with(roc_df, {
  lines(ffp, fvp, type='l', lwd=1, col="blue")
  lines(ffp, fvp, type='b', lwd=1, col="blue")
})
```

El dibujo automático de la curva *roc* (opción 1 del código en R) se muestra en la figura 2.1. Hay que observar que R etiqueta el eje de abscisas al revés (de derecha a izquierda). Por otra parte, dado que el biomarcador X no es bueno, su rendimiento está por debajo de la diagonal, R interpreta que X es un anti-biomarcador con lo que cambia el $\widehat{auc}$ al complementario, proporcionando $1 - 0.4143 = 0.5857$.

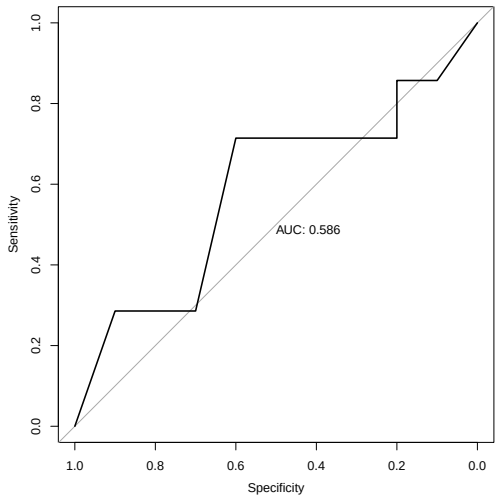

Figura 2.1: Curva *roc* automática para los datos observados del esquema $D \leftarrow C$

Es posible construir una curva *roc* personalizada según la figura 2.2 (opción 2 del código en R) en la que el eje de abscisas esté etiquetado de la forma habitual y que incluya una retícula que ayude a comprender las áreas trapezoidales.

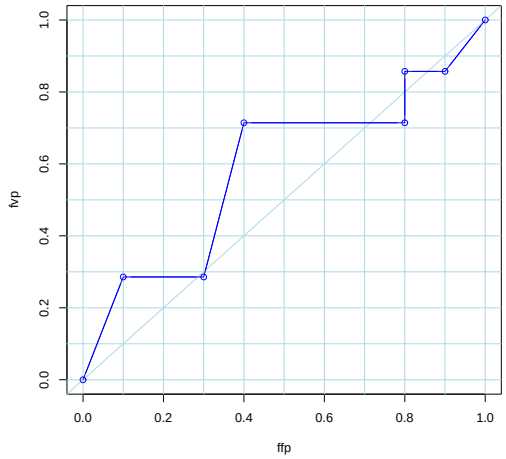

Figura 2.2: Curva *roc* personalizada para los datos observados del esquema $D \leftarrow C$

Referencias

[1] Davies, T.M., *The Book of R: A first course in programming and statistics*, No Starch Press, 2016.

[2] García-Pérez, A., *Estadística básica con R*, Universidad Nacional de Educación a Distancia (UNED) (Madrid), 2010.

[3] Letón, E., Pedromingo, A., *Introducción al análisis de datos en meta-análisis*, Díaz de Santos (Madrid), 2001.

[4] Peña, D., *Fundamentos de Estadística*, Alianza editorial (Madrid), 2014.

[5] Pepe, M.S., *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, 2003.

[6] R Core Team, *R: A Language and Environment for Statistical Computing. R*, Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[7] Ramos, E., Vélez, R., Hernández, V., *Modelos probabilísticos y optimización*, Sanz y Torres (Madrid), 2019.

[8] Redondo, F.L., *La lógica en la interpretación de las pruebas diagnósticas*, Garsi (Barcelona), 1989.

Acerca de

Emilio Letón

Figura 2.3: Emilio Letón

Emilio Letón nace en Madrid en 1966. Es Licenciado en Matemáticas por la UCM en 1989 y doctor en Matemáticas en la misma universidad en 2002. En la actualidad es profesor contratado doctor de la UNED en el departamento de Inteligencia Artificial, al que se incorporó en 2009. Anteriormente fue profesor del departamento de Estadística de la UC3M durante 5 años. Asimismo, ha trabajado durante 15 años en departamentos de Planificación y Estadística dentro del sector bancario y de la industria farmacéutica. Sus líneas de investigación incluyen el Análisis de Supervivencia, tests no paramétricos, PLS, Meta-Análisis, Bioestadística y B-Learning. Ha participado en más de 30 proyectos de innovación docente (siendo coordinador en más de 10 de ellos) colaborando con distintas universidades: UNED, UC3M, UCM y UPM. Ha recibido 1 premio en excelencia en publicaciones científicas (UC3M) y 5 premios en excelencia docente (1 en UC3M, 3 en UNED y 1 en OCW Consortium). En @emilioleton se pueden encontrar sus tweets y su página web personal con información ampliada de su curriculum. https://twitter.com/emilioleton

Elisa M. Molanes-López

Figura 2.4: Elisa M. Molanes-López

Elisa M. Molanes López nace en Vigo en 1976. Se licencia en Matemáticas por la Universidad de Santiago de Compostela en 2000 y consigue el grado de doctora, con acreditación europea, por la Universidad de A Coruña en 2007. Entre 2003 y 2007, al amparo de una beca FPI, realiza varias estancias de investigación

en las siguientes universidades extranjeras: Universiteit Hasselt (Bélgica), Université Catholique de Louvain (Bélgica) y The University of Texas (EE.UU.). Durante 8 años, entre 2007 y 2015, es profesora visitante en el departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente, desde octubre de 2015, es profesora ayudante doctora en la Unidad Departamental de Bioestadística del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid. Durante su etapa docente ha participado en 10 proyectos de innovación docente. En diciembre de 2013, recibe un accésit a la mejor práctica docente en los Premios del Consejo Social de la UNED por su participación en el MOOC "Mini-vídeos docentes modulares: un elemento crítico en el diseño de un MOOC", y en abril de 2014, el OpenCourseWare Consortium le otorga un premio de excelencia por su participación en el curso OCW "Mini-vídeos docentes modulares para diseñar un MOOC". Sus líneas de investigación incluyen la estadística no paramétrica, el análisis de supervivencia, las curvas ROC y las funciones cópula. En la página web personal de Elisa M. Molanes-López, se puede encontrar información más detallada. https://elisamariamolanes.wixsite.com/emolanes

Índice general

Índice de figuras

Índice de tablas