Curvas R.O.C.

Carlos Carleos Norberto Corral

Departamento de Estadística
e Investigación Operativa
y de Didáctica de la Matemática
Universidad de Oviedo

8 de noviembre de 2021

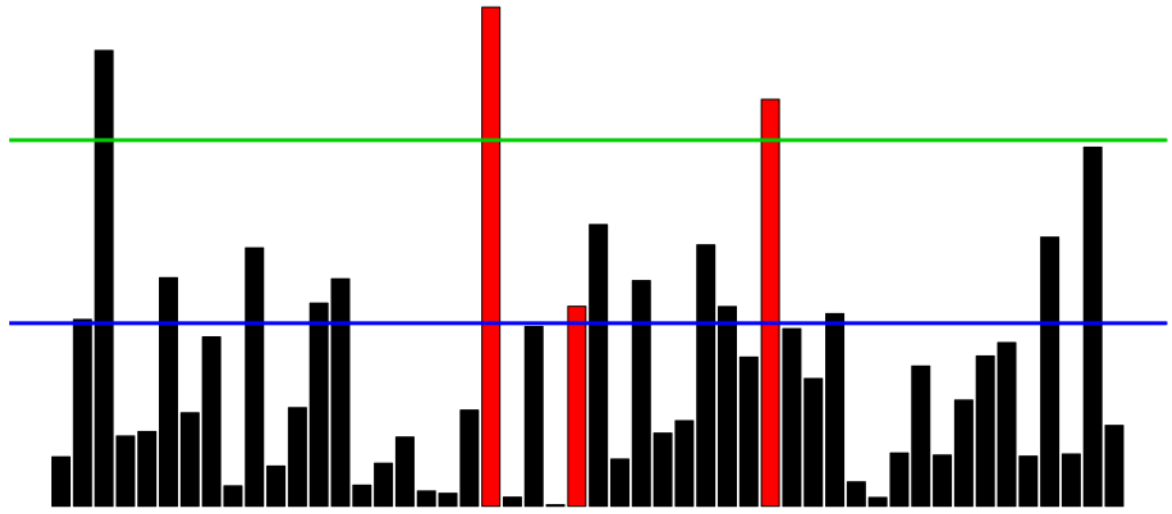
Origen

Receiver Operating Characteristic (ROC)

- Proviene de la teoría de detección de señales (discernir entre señal y ruido)
- Gráfico estadístico
- Ayuda a definir un umbral de separación entre dos grupos
  - Positivos (los que verifican cierta condición)
  - Negativos (los que no la cumplen)

Ejemplo

Señal en radar: ¿ruido o aviones?



Umbral de detección (decidir entre “ruido” o “avión”)

- umbral alto
  no se detectan objetos reales (Falsos Negativos)
- umbral bajo
  muchos avisos falsos (Falsos Positivos)

Estado

El estado indica el grupo al que realmente pertenece cada individuo.

- Variable aleatoria D que toma los valores

$$D = \begin{cases} 0 & \text{No verifica la condición (Negativo)} \\ 1 & \text{Si verifica la condición (Positivo)} \end{cases}$$

- $D \rightsquigarrow \text{Bernoulli}(p)$

$$\begin{aligned} p &= \Pr(D = 1) = \Pr(\text{Positivo}) = \text{prevalencia} \\ 1 - p &= \Pr(D = 0) = \Pr(\text{Negativo}) \end{aligned}$$

Variable medida

- X representa a la variable que se usa para establecer el criterio de clasificación
  - En los Positivos $X = X_P = (X \mid D = 1)$
  - En los Negativos $X = X_N = (X \mid D = 0)$
- Sus correspondientes funciones de distribución son
  - $F_P(x) = \Pr(X_P \leq x)$
  - $F_N(x) = \Pr(X_N \leq x)$

Variable criterio de clasificación

Cuando X tiende a tomar valores mayores para los individuos positivos la clasificación se suele asociar a alcanzar cierto umbral c:

$$Y = \begin{cases} 1 \text{ (Positivo)} & \text{si } X \geq c \\ 0 \text{ (Negativo)} & \text{si } X < c \end{cases}$$

Estimador de la prevalencia

- Proporción muestral

$$p = \frac{\text{número de individuos Positivos de la muestra}}{\text{número de individuos de la muestra}}$$

- En algunos tipos de muestreo, por ejemplo caso-control, no es una buena estiamción.

Términos habituales

VP	verdadero positivo	$\Longleftrightarrow$	$D = 1, Y = 1$
FP	falso positivo	$\Longleftrightarrow$	$D = 0, Y = 1$
FN	falso negativo	$\Longleftrightarrow$	$D = 1, Y = 0$
VN	verdadero negativo	$\Longleftrightarrow$	$D = 0, Y = 0$

Sensibilidad

- Probabilidad de acertar al clasificar un individuo positivo

$$S = \Pr(Y = 1 \mid D = 1) = \frac{\Pr(Y = 1 \cap D = 1)}{\Pr(D = 1)}$$

$$S = \Pr(X_p \geq c)$$

$$\hat{S} = \frac{VP}{VP + FN} = \text{fracción de verdaderos positivos} = FVP$$

$$1 - \hat{S} = \frac{FN}{VP + FN} = \text{fracción de falsos negativos} = FFN$$

Especificidad

- Probabilidad de acertar al clasificar un individuo negativo

$$E = \Pr(Y = 0 \mid D = 0) = \frac{\Pr(Y = 0 \cap D = 0)}{\Pr(D = 0)}$$

$$E = \Pr(X_n < c)$$

$$\hat{E} = \frac{VN}{VN + FP} = \text{fracción de verdaderos negativos} = FVN$$

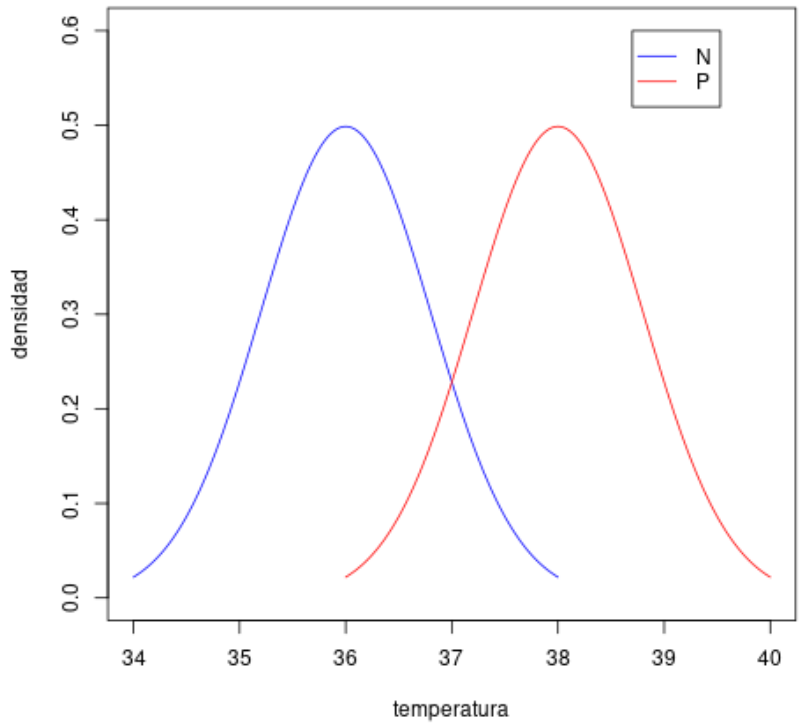
$$1 - \hat{E} = \frac{FP}{VN + FP} = \text{fracción de falsos positivos} = FFP$$

Valores predictivos

- Valor predictivo positivo (Probabilidad a posteriori)
  - $\Pr(D = 1 \mid Y = 1)$
  - Se estima mediante $\widehat{VPP} = \dfrac{VP}{VP + FP}$
- Valor predictivo negativo (Probabilidad a posteriori)
  - $\Pr(D = 0 \mid Y = 0)$
  - Se estima mediante $\widehat{VPN} = \dfrac{VN}{VN + FN}$

Definición de curva ROC

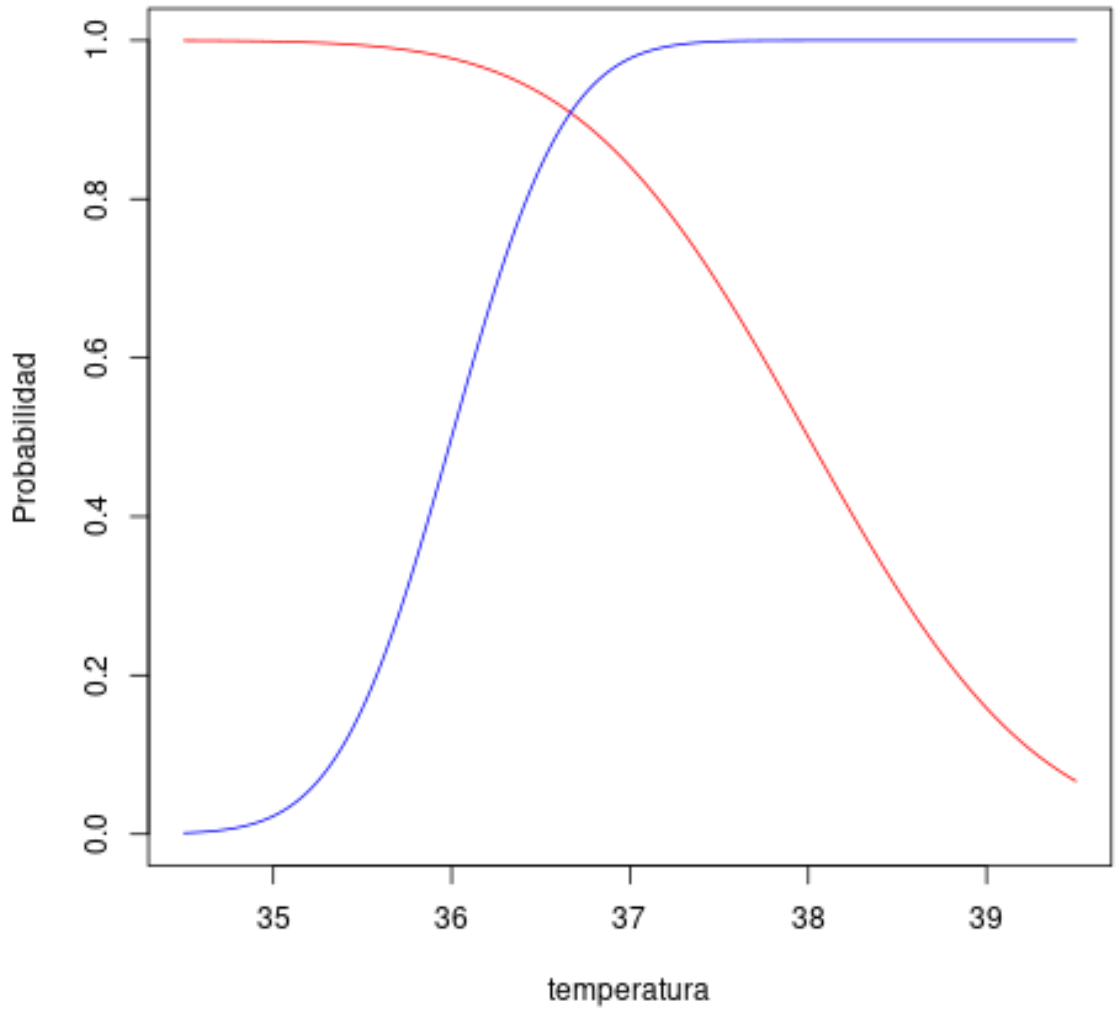
Ejemplo: Diagnosticar una enfermedad teniendo en cuenta la temperatura.



$$Y = \begin{cases} 1 \text{ (Positivo)} & \text{temperatura} \geq c \\ 0 \text{ (Negativo)} & \text{temperatura} < c \end{cases}$$

Definición de curva ROC

Objetivo: Elegir *c* que tenga la Sensibilidad y la Especificidad altas.



Definición de curva ROC

La curva ROC representa, para cada valor del umbral la Sensibilidad frente a 1-Especificidad, es decir,

- Abscisas: 1-Especificidad= Pr(Error|D=0)
- Ordenadas: Sensibilidad= Pr(Acierto|D=1)

Suponiendo que X toma, en general, valores mayores para los positivos y un umbral x, se tiene:

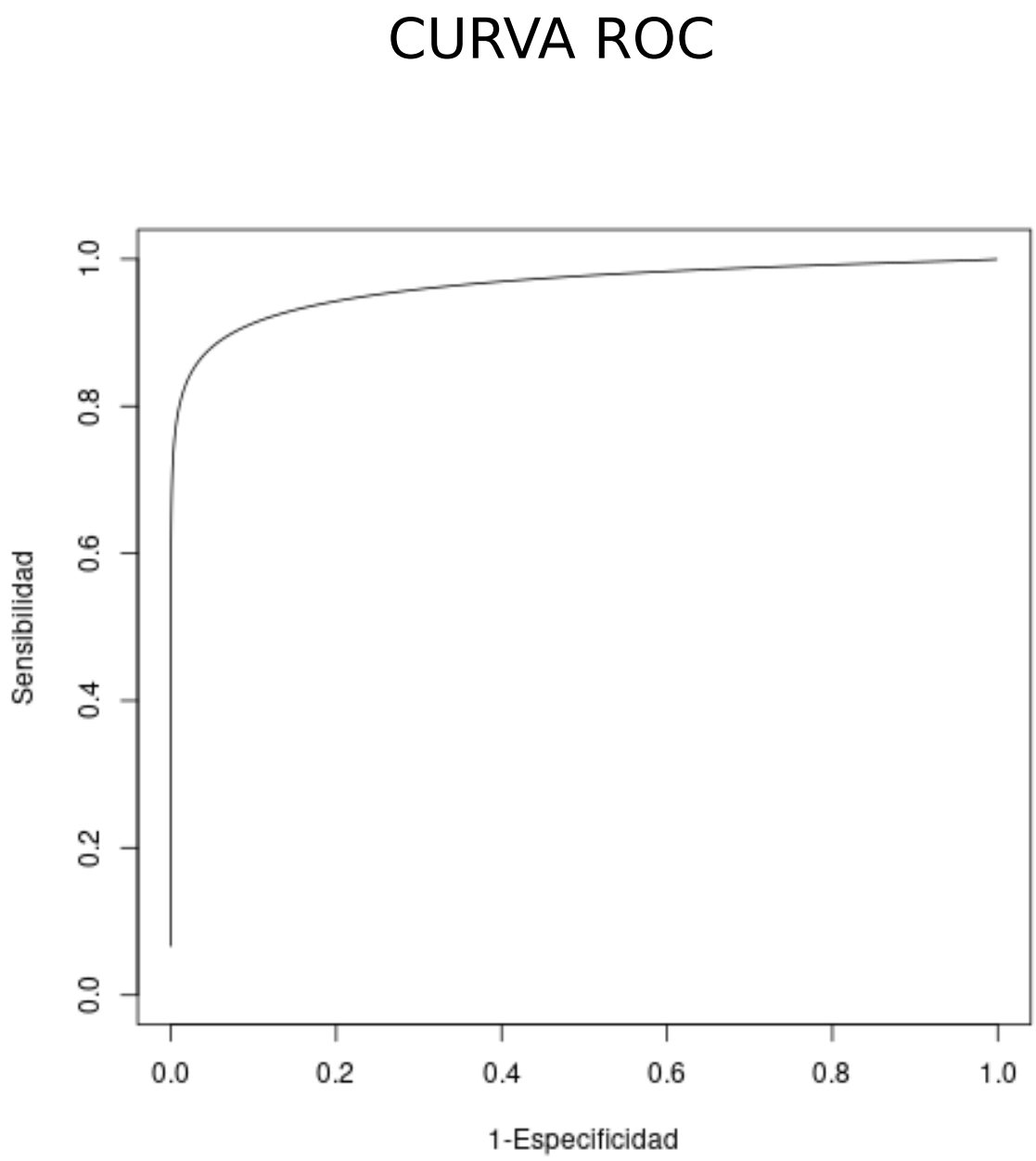
$$t = 1 - E(x) = \Pr\left(Y(x) = 1 \mid D = 0\right) = 1 - F_N(x)$$

$$x = F_N^{-1}(1-t)$$

$$ROC(t) = Pr\left(\text{Acierto} \mid \text{Positivo }\right) = S(x) = 1 - F_P(x)$$

$$ROC\left(t\right) = 1 - F_P\left(F_N^{-1}(1-t)\right) \qquad 0 \leq t \leq 1$$

Definición de curva ROC



Método no paramétrico para estimar la curva ROC

- Usa la función de distribución empírica asociada a la muestra:

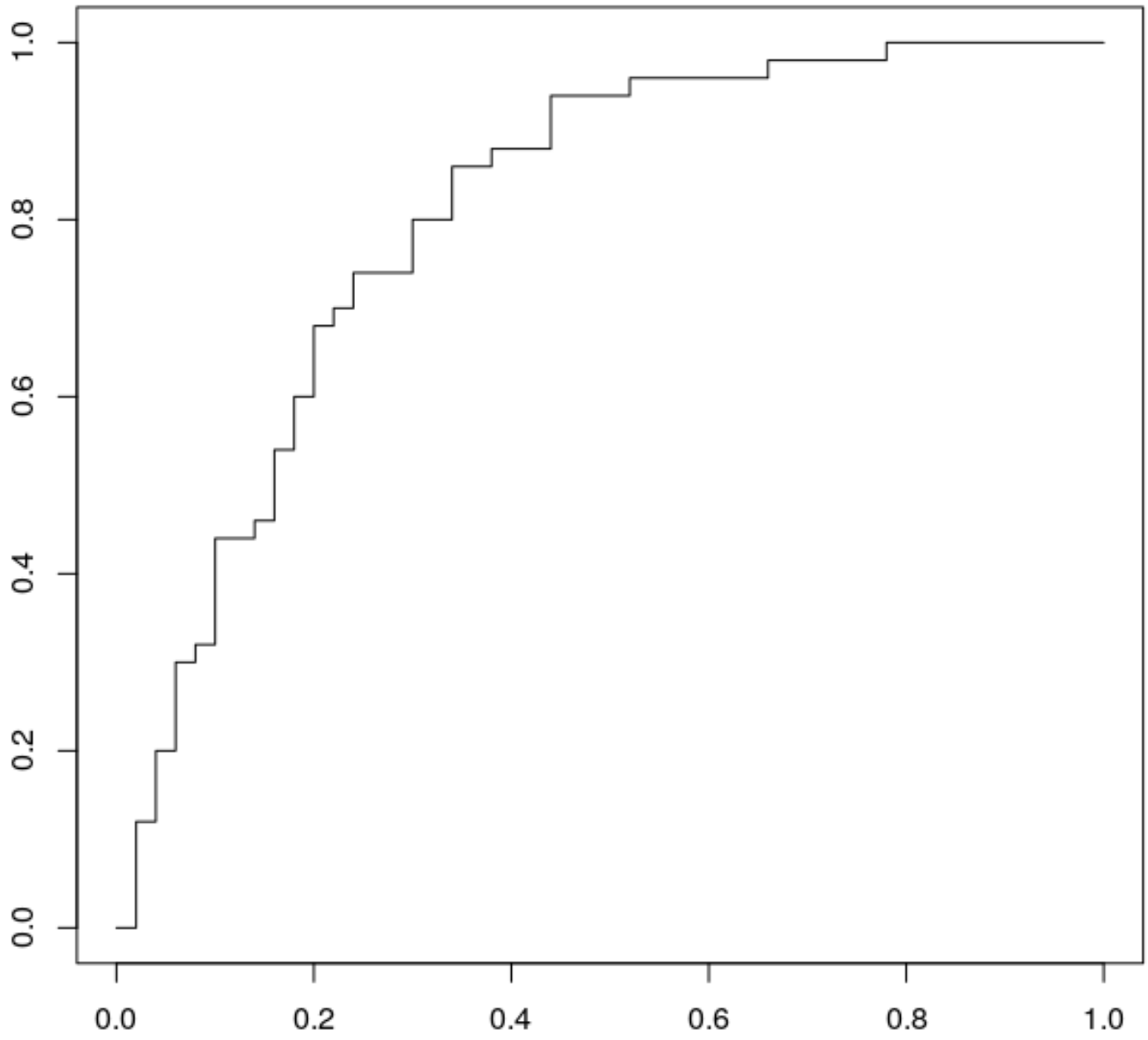
$$\hat{F}_P(x) = \frac{n(x_i \leq x \mid Positivos)}{n_P}$$
$$\hat{F}_N(x) = \frac{n(x_i \leq x \mid Negativos)}{n_N}$$

- Se define la curva ROC empírica como

$$\widehat{ROC}(t) = 1 - \hat{F}_P\left(\hat{F}_N^{-1}(1-t)\right)$$

- Este procedimiento da lugar a una curva ROC escalonada

Método no paramétrico para estimar la curva ROC



Método paramétrico para estimar la curva ROC

- Modelo binormal
  - $(X_P \rightsquigarrow N(\mu_P, \sigma_P)$ Positivos
  - $X_N \rightsquigarrow N(\mu_N, \sigma_N)$ Negativos
- La curva ROC en el modelo binormal es

$$ROC(t) = 1 - \Phi\left(\frac{\mu_N - \mu_P + \sigma_N \cdot \Phi^{-1}(1-t)}{\sigma_P}\right) \qquad 0 \leq t \leq 1$$

$$= \Phi\left(\frac{\mu_P - \mu_N + \sigma_N \cdot \Phi^{-1}(t)}{\sigma_P}\right) = \Phi\left(\alpha + \beta \cdot \Phi^{-1}(t)\right)$$

donde $\alpha = \dfrac{\mu_P - \mu_N}{\sigma_P}$, $\beta = \frac{\sigma_N}{\sigma_P}$,

Φ = función de distribución de la gausiana típica

Método paramétrico para estimar la curva ROC

- La curva ROC estimada es

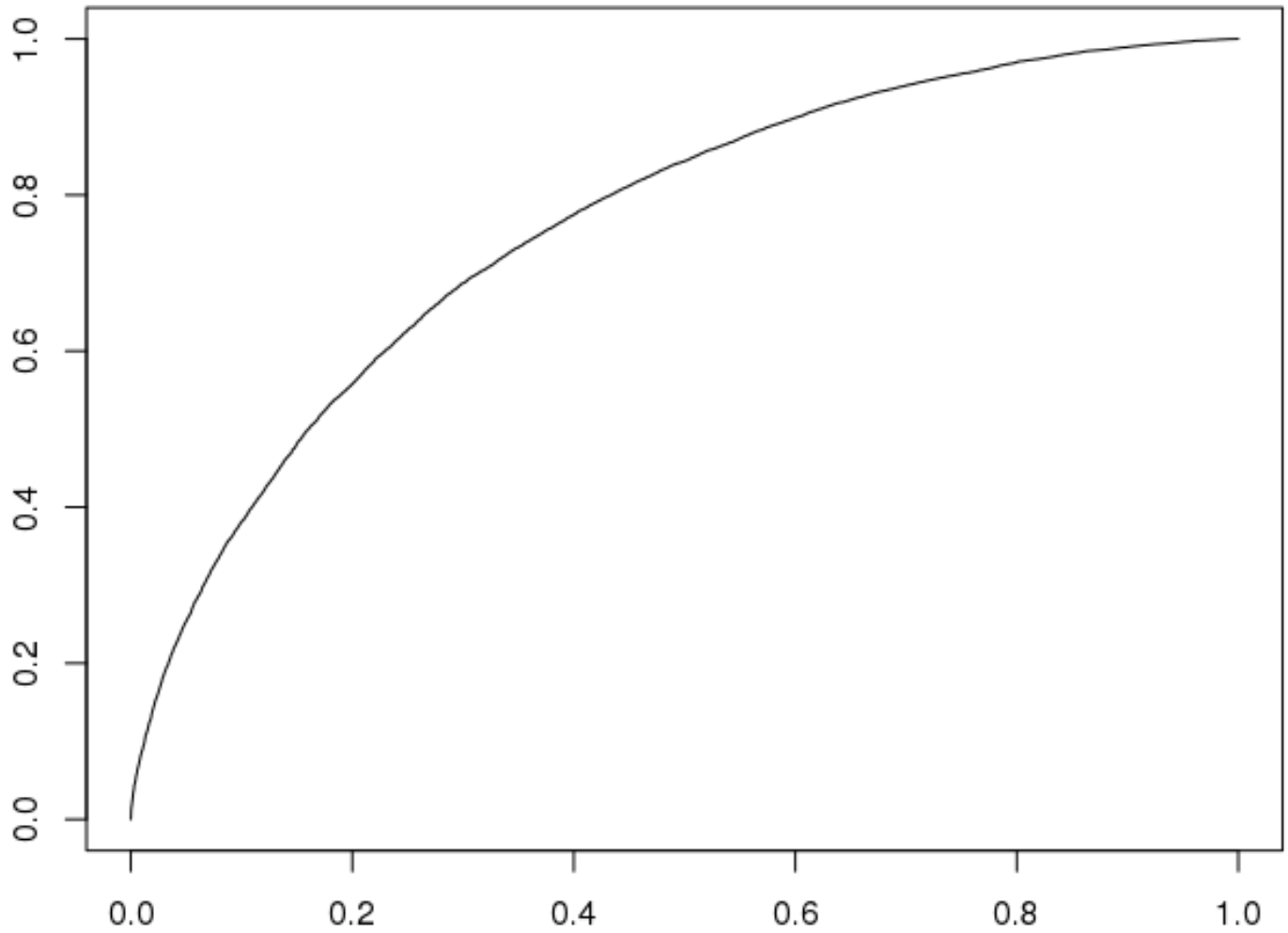
$$\widehat{ROC}(t) = 1 - \Phi\left(\hat{\alpha} + \hat{\beta} \cdot \Phi^{-1}(t)\right) \qquad 0 \leq t \leq 1$$

donde

$$\hat{\alpha} = \frac{\hat{\mu}_P - \hat{\mu}_N}{\hat{\sigma}_P} \qquad \hat{\beta} = \frac{\hat{\sigma}_N}{\hat{\sigma}_P}$$

- Es una curva suave y derivable en todos los puntos.

Método paramétrico para estimar la curva ROC



Precisión o exactitud de un clasificador

$$\begin{aligned}\Pr(\text{acertar}) &= \Pr(Y=1 \mid D=1)\cdot\Pr(D=1)\\ &+ \Pr(Y=0 \mid D=0)\cdot\Pr(D=0)\\ &= S\cdot\Pr(D=1)+E\cdot\Pr(D=0)\end{aligned}$$

- Estimador
$$\widehat{\Pr}(acertar)=\frac{VP+VN}{n}=\frac{\text{resultados acertados}}{\text{total de la muestra}}$$
- Proporción de aciertos en la clasificación sin distinguir positivos de negativos
- Es necesario conocer o estimar la prevalencia $Pr(D=1)$

Área bajo la curva

El Área Bajo la Curva AUC (*area under curve*) se define así:

$$AUC = \int_0^1 ROC(t)\,\mathrm{d}t$$

$AUC = P(X_P > X_N)$

Demostración:
$AUC = \int_0^1 ROC(t)\,\mathrm{d}t = \int_0^1 \left[1 - F_p\left(F_N^{-1}(1-t)\right)\right]\mathrm{d}t$
Haciendo el cambio de variable $u = F_N^{-1}(1-t)$ se tiene $F_N(u) = 1 - t \iff t = 1 - F_N(u)$.
Por tanto $\mathrm{d}t = -f_N(u)\mathrm{d}u$ y la integral anterior es
$\int_{-\infty}^{\infty}(1 - F_p(u))f_N(u)\,\mathrm{d}u = 1 - \int_{-\infty}^{\infty} F_p(u)f_N(u)\,\mathrm{d}u =$
$1 - \int_{-\infty}^{\infty}\left(\int_{-\infty}^{v} f_p(v)\mathrm{d}v\right)f_N(u)\,\mathrm{d}u = 1 - P(X_p \leq X_N) =$
$P(X_p > X_N)$

Área bajo la curva

Propiedades del Área Bajo la Curva AUC:

- $AUC = P(X_P > X_N)$
- $0{,}5 \leq AUC \leq 1$
- $AUC \approx 0{,}5 \iff$ poca capacidad de discriminación
- $AUC \approx 1 \iff$ separación casi total
- Se usa con mucha frecuencia para medir la capacidad de una variable para separar dos poblaciones

Métodos para calcular el AUC

- Método no paramétrico (regla trapezoidal)

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

- Método paramétrico (caso binormal)

$$\begin{aligned} \widehat{AUC} &= \int_0^1 \widehat{ROC}(t)\,\mathrm{d}t = \int_0^1 \Big(1 - \Phi\big[\hat{\alpha} + \hat{\beta} \cdot \Phi^{-1}(1-t)\big]\Big)\mathrm{d}t \\ &= \Phi\left(\frac{\hat{\alpha}}{\sqrt{1+\hat{\beta}^2}}\right) = \Phi\left(\frac{\hat{\mu}_N - \hat{\mu}_P}{\sqrt{\hat{\sigma}_N^2 + \hat{\sigma}_P^2}}\right) \end{aligned}$$

Comparación de dos curvas ROC

- Contraste

$$H_0: \quad AUC_1 = AUC_2$$

$$H_1: \quad AUC_1 \neq AUC_2$$

- Estadístico

$$z = \frac{\widehat{AUC}_1 - \widehat{AUC}_2}{ET(\widehat{AUC}_1 - \widehat{AUC}_2)}$$

Criterios para elegir el punto de corte

- Método 1. El índice de Youden
- Método 2. Buscar el punto (1-E;S) sobre la curva ROC más cercano al punto (0;1).

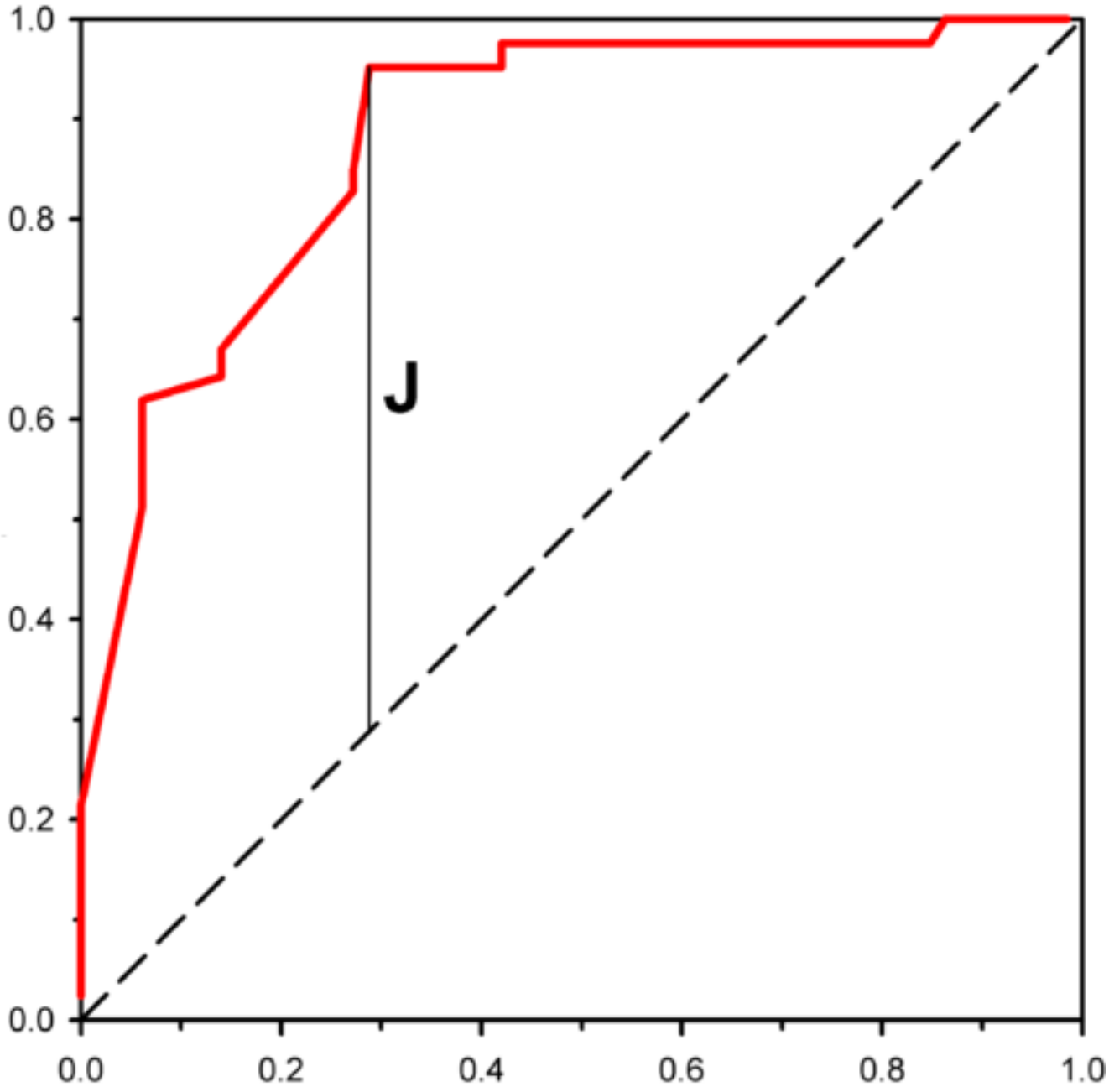
Índice de Youden

- Busca el umbral que hace máxima la suma de la sensibilidad y la especificidad

$$\begin{aligned} J &= \Pr(Y = 1 \mid D = 1) + \Pr(Y = 0 \mid D = 0) - 1 \\ &= \text{Sensibilidad} + \text{Especificidad} - 1 \end{aligned}$$

- $0 \leq J \leq 1$
- $J \approx 0 \iff$ discrimina poco entre positivos y negativos
- $J \approx 1 \iff$ discrimina mucho entre positivos y negativos
- Su estimador se define como $\hat{J} = FVP - FFP$

Índice de Youden



Punto de corte: método 2

- Buscar el punto (1-E;S) sobre la curva ROC más cercano al punto (0;1).
  El umbral que se elige está asociado al punto sobre la curva que hace mínima la siguiente expresión:

$$(1 - Especificidad)^2 + (1 - Sensibilidad)^2$$

Biblioteca «pROC»

```
nnN <- nP <- 50 # tamaño de la muestra
# simulación de los datos de la variable X1
N1 <- rnorm (nN, 36, 0.8) # negativos
P1 <- rnorm (nP, 38, 0.8) # positivos
# install.packages ("pROC") # instalar la librería
library (pROC) # carga la librería pROC)
C1 <- roc(controls=N1,cases=P1) # curva ROC C1
plot (C1, legacy.axes=TRUE) # gráfica de la curva
auc (C1)     # área bajo la curva
ci(C1) ## Intervalo confianza AUC
```

Biblioteca «pROC»

```
# criterios para elegir el umbral
## método 1: Youden
coords(C1,"best")
# método 2
coords(C1,"best",best.method="closest.topleft")

# simulación de los datos de la variable X2
N2 <- rnorm (nN, 36.5, 0.8) # negativos
P2 <- rnorm (nP, 37.5, 0.8) # positivos
C2 <- roc(controls=N2,cases=P2) # curva ROC 2

# gráfica con las dos curvas
plot (C2, add=TRUE, col="red")

roc.test (C1, C2) # compara las curvas C1 y C2
```