CORvas

Curvas R.O.C.

Carlos Carleos Norberto Corral

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

Máster de Análisis de Datos
para Inteligencia de Negocios

Partes

Teoría

- Motivación, conceptos previos y definición
- Inferencia
- Elección del umbral

Práctica en R

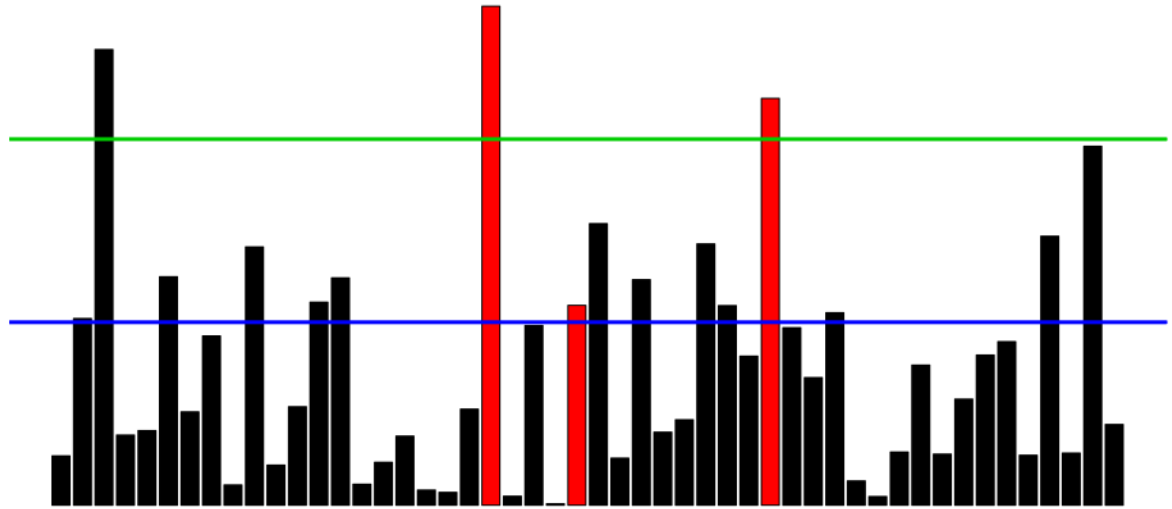
Origen

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

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

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

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

Variable medida

- Variable X: 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 queda determinada dado 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, como el caso-control, puede dar lugar a estimaciones muy sesgadas.

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

$$S = \Pr(X_P \geq c) = \Pr(Y = 1 \mid D = 1)$$

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

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

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

Especificidad

- Probabilidad de acertar al clasificar un individuo negativo

$$E = \Pr(X_N < c) = \Pr(Y = 0 \mid D = 0)$$

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

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

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

Fracciones (tabla resumen)

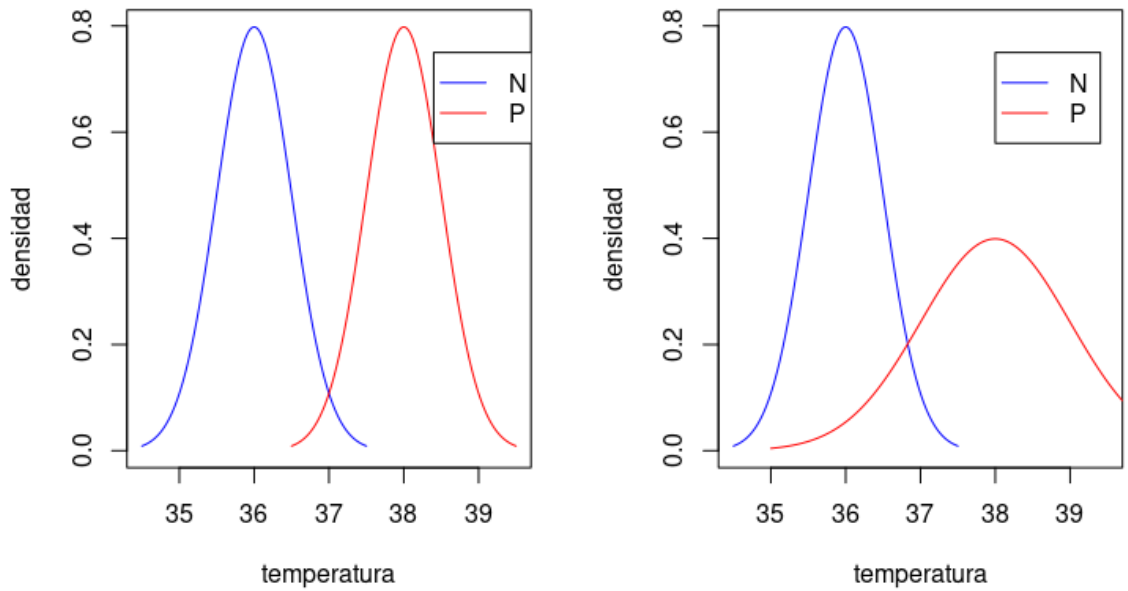
	$Y = 1$	$Y = 0$
$D = 1$	FVP	FFN
$D = 0$	FFP	FVN

Valores predictivos

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

Definición de curva ROC

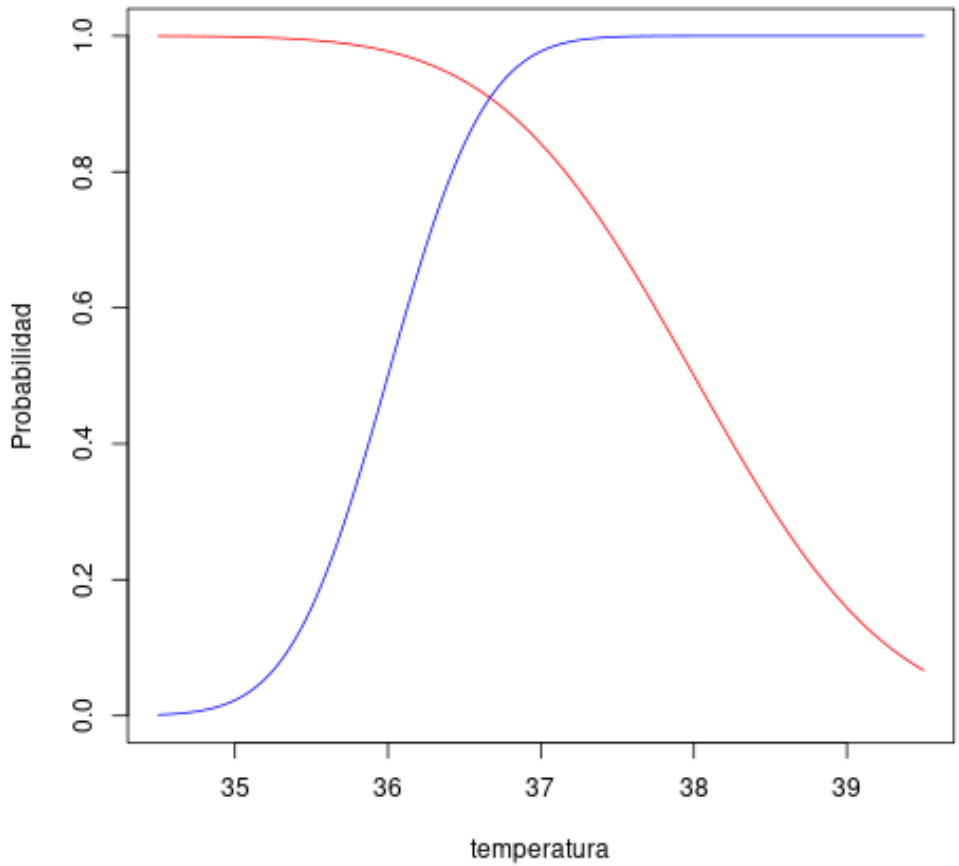
Objetivo: Seleccionar el umbral de temperatura c que optimice la decisión.



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

Definición de curva ROC

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



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

Definición de curva ROC

- Curva ROC **poblacional**
- Gráfica que representa, para cada valor del umbral c, la Sensibilidad frente a $1 -$ Especificidad, es decir,
  - Abscisas $= 1 -$ Especificidad $= \Pr(\text{Error} \mid D = 0)$
  - Ordenadas $=$ Sensibilidad$= \Pr(\text{Acierto} \mid D = 1)$

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

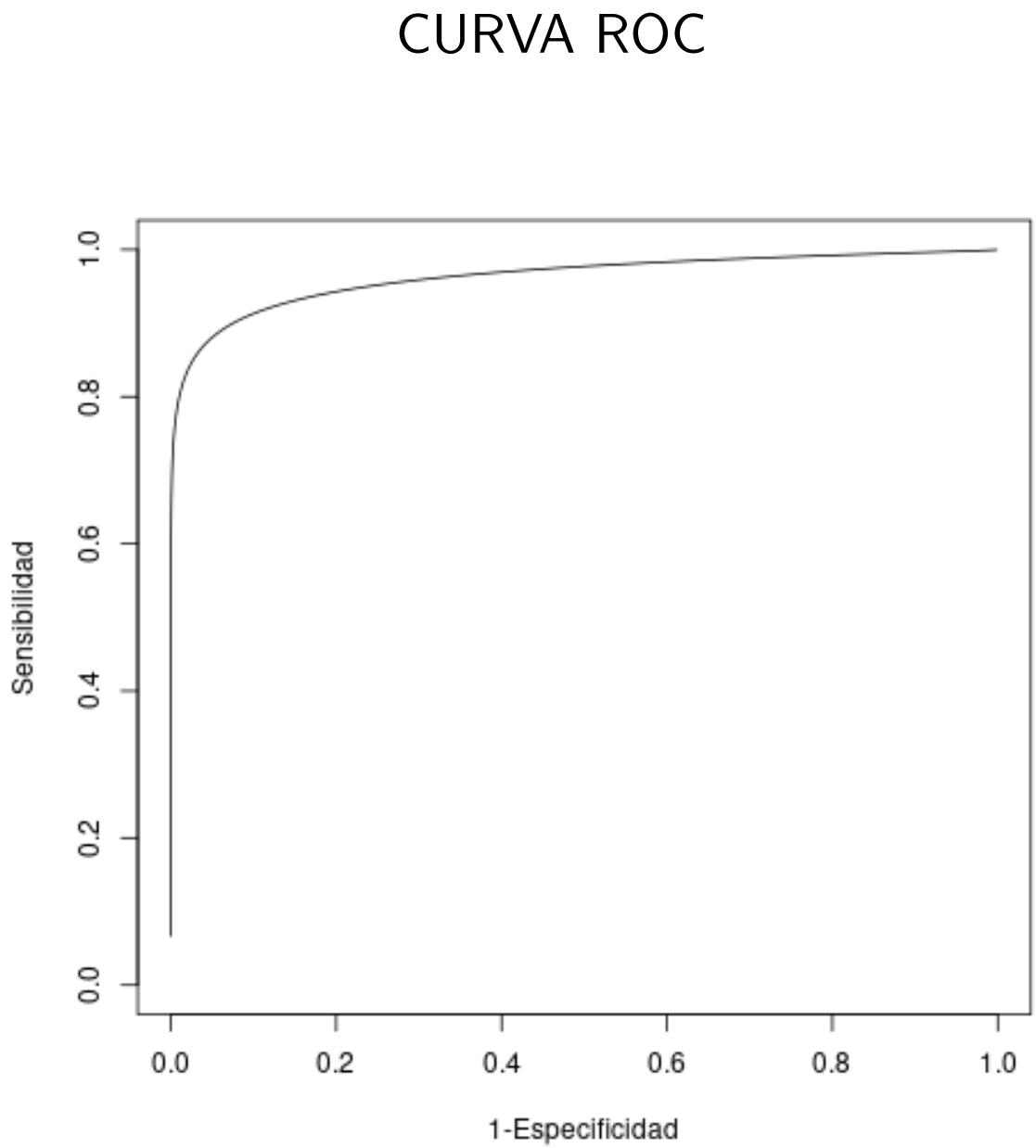
$$x = F_N^{-1}(1 - t)$$

- Suponiendo que X toma, en general, valores mayores para los positivos,

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

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

Definición de curva ROC



Método no paramétrico para estimar la curva ROC

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

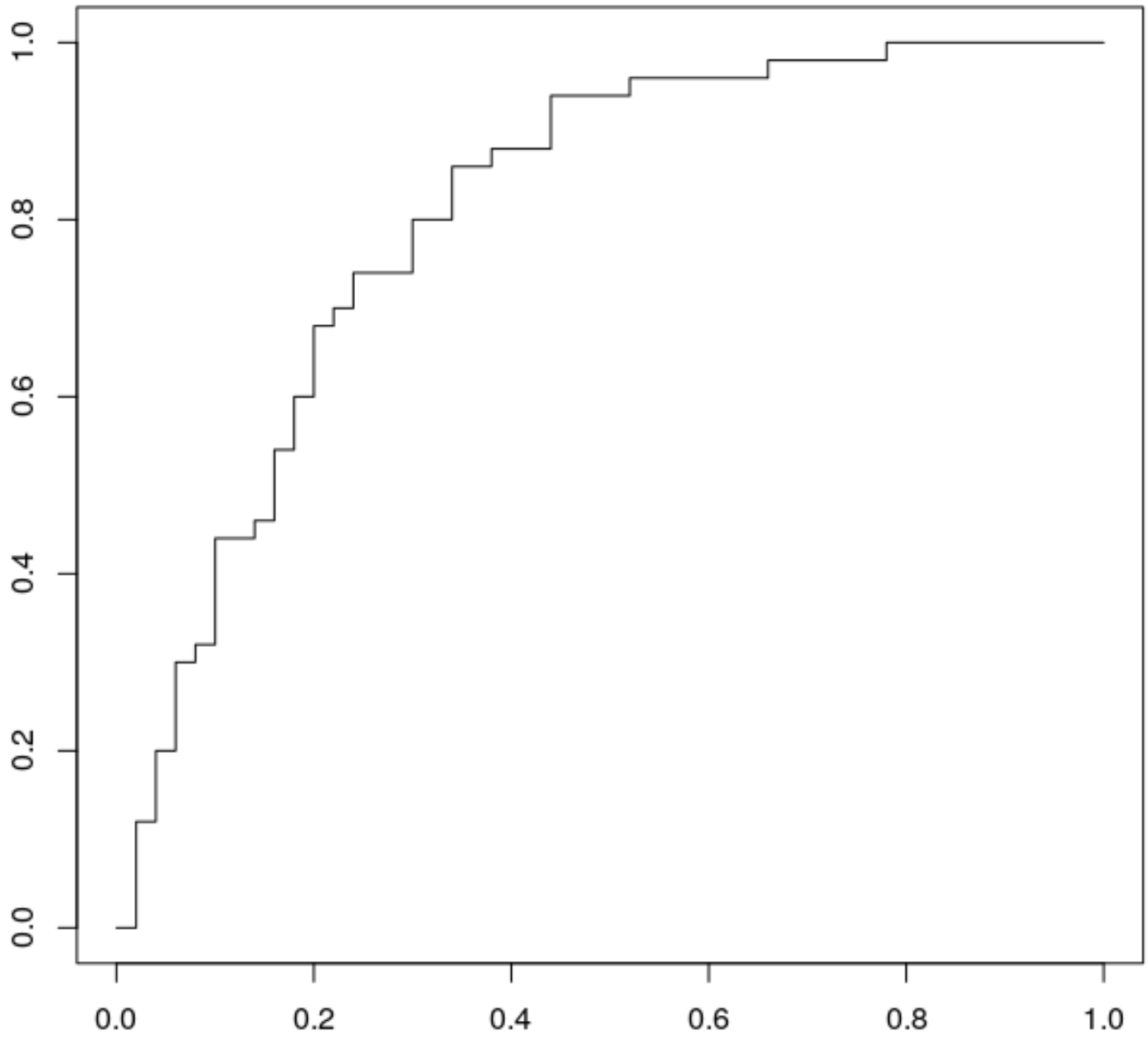
$$\hat{F}_P(x) = \frac{\text{n}^{\underline{o}}\,(x_i \leq x \mid \text{Positivos})}{n_P}$$
$$\hat{F}_N(x) = \frac{\text{n}^{\underline{o}}\,(x_i \leq x \mid \text{Negativos})}{n_N}$$

- Se define la curva ROC empírica como

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

- 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

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

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

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

$\Phi =$ función de distribución de la gausiana típica

Método paramétrico para estimar la curva ROC

- La curva ROC estimada es

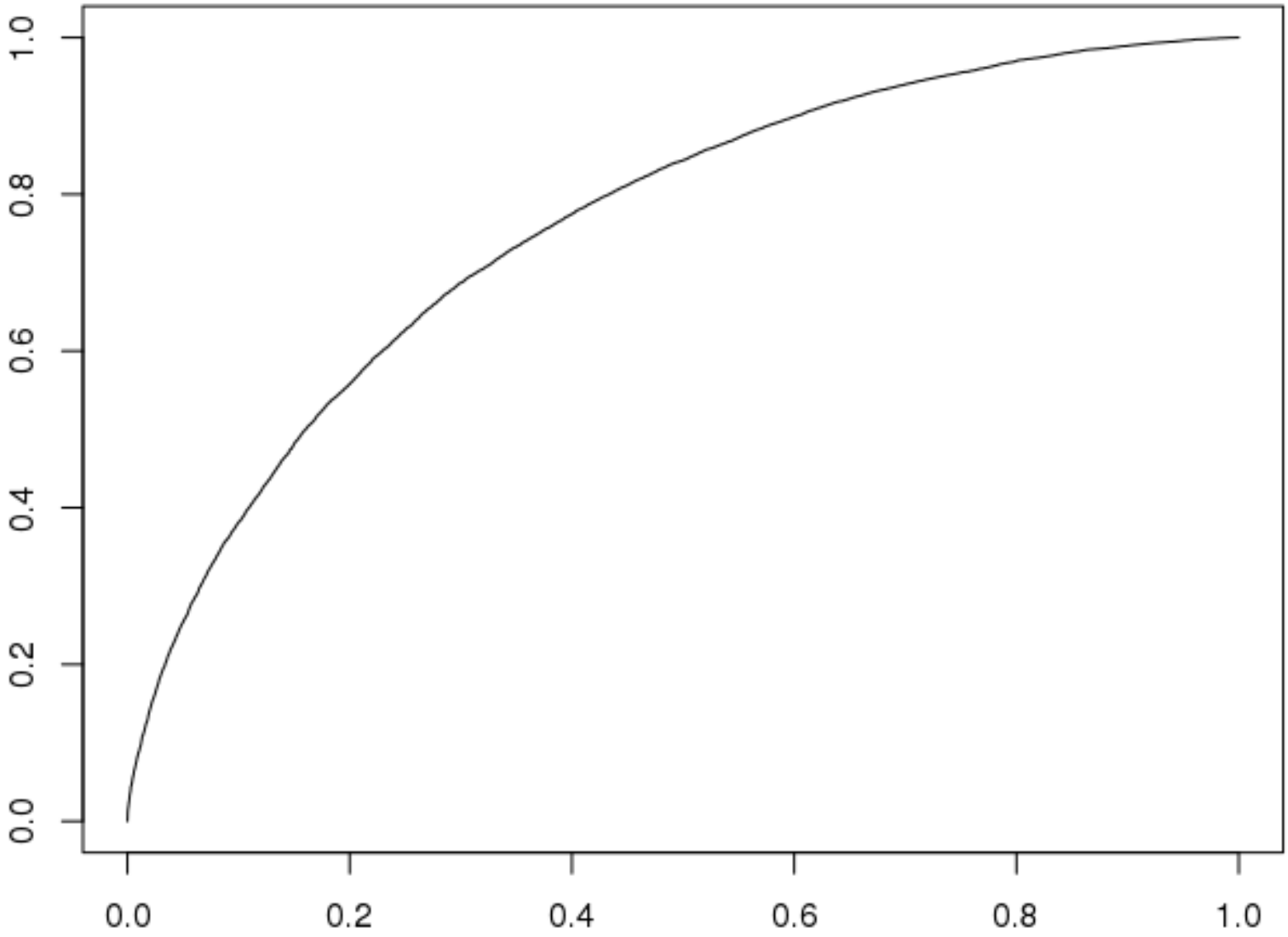
$$\widehat{\text{ROC}}(t) = 1 - \Phi\left(\hat{\alpha} + \hat{\beta} \cdot \Phi^{-1}(t)\right) \qquad 0 \leq t \leq 1$$

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



Contrastes de hipótesis

- En el caso concreto del modelo binormal

$$H_0: \quad \alpha_1 = \alpha_2 \quad \cap \quad \beta_1 = \beta_2$$

$$H_1: \quad \alpha_1 \neq \alpha_2 \quad \cup \quad \beta_1 \neq \beta_2$$

- El estadístico del contraste:

$$\chi^2 = \frac{\hat{\alpha}_{12}^2 \cdot \hat{V}(\hat{\beta}_{12}) + \hat{\beta}_{12}^2 \cdot \hat{V}(\hat{\alpha}_{12}) - 2 \cdot \hat{\alpha}_{12} \cdot \hat{\beta}_{12} \cdot \widehat{\mathrm{cov}}(\hat{\alpha}_{12}, \hat{\beta}_{12})}{\hat{V}(\hat{\alpha}_{12}) \cdot \hat{V}(\hat{\beta}_{12}) - \widehat{\mathrm{cov}}(\hat{\alpha}_{12}, \hat{\beta}_{12})}$$

donde $\hat{\alpha}_{12} = \hat{\alpha}_1 - \hat{\alpha}_2$ y $\hat{\beta}_{12} = \hat{\beta}_1 - \hat{\beta}_2$

- Asintóticamente bajo H_0, $\chi^2 \rightsquigarrow \chi_2^2$

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

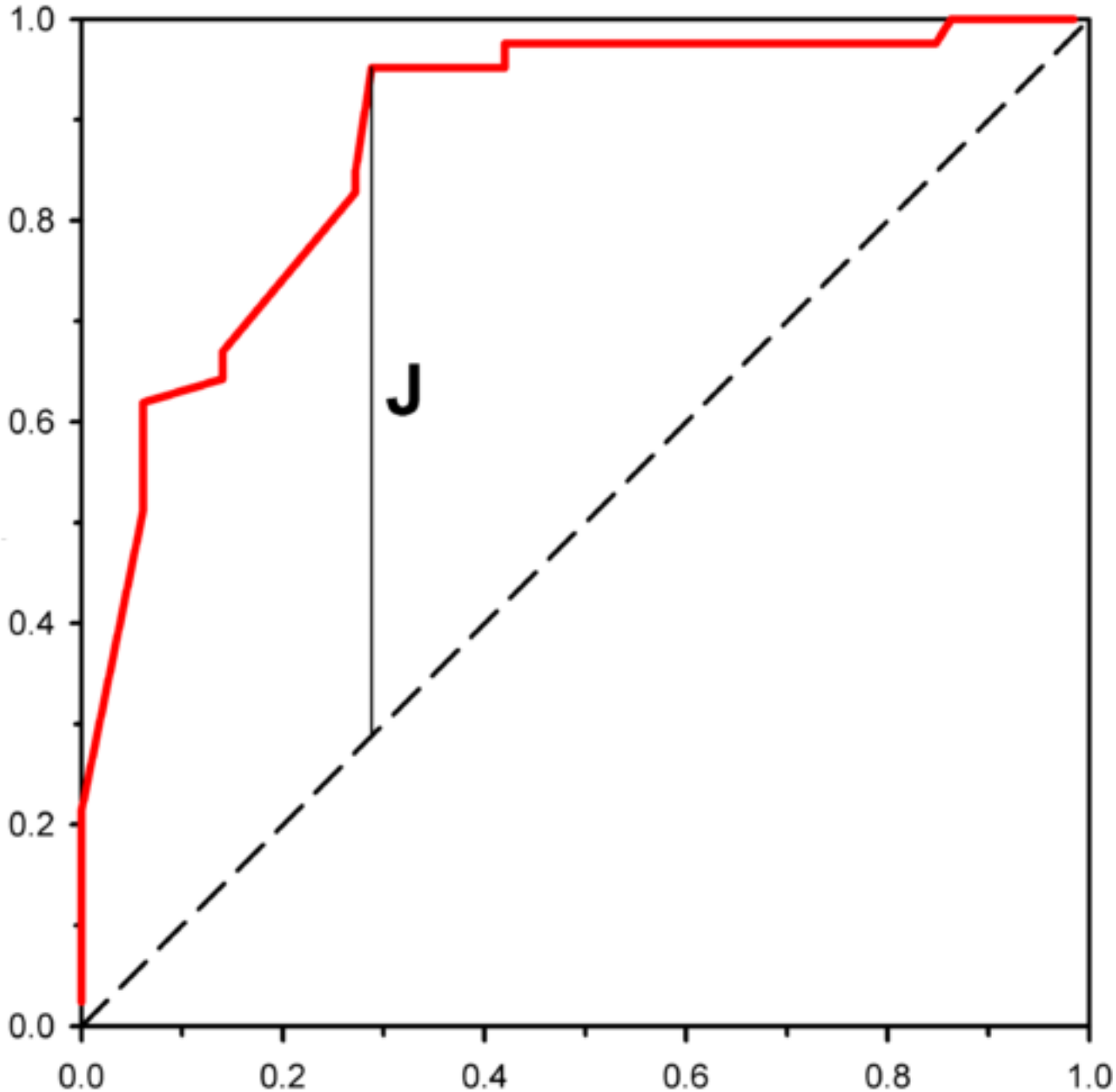
Índice de Youden

- Es la diferencia entre las probabilidades de la respuesta positiva correcta y de la respuesta positiva incorrecta

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

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

Índice de Youden



Área bajo la curva

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

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

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

Área bajo la curva

$$\text{AUC} = \Pr\left(X_P > X_N\right)$$

Demostración:

$$\text{AUC} = \int_0^1 \text{ROC}(t)\,\mathrm{d}t = \int_0^1 \left[1 - F_P\left(F_N^{-1}(1-t)\right)\right]\,\mathrm{d}t$$

Haciendo el cambio de variable $u = F_N^{-1}(1-t)$ se tiene

$$F_N(u) = 1 - t \Longleftrightarrow t = 1 - F_N(u)$$

Por tanto $\mathrm{d}t = -f_N(u)\,\mathrm{d}u$ y la integral anterior es

$$\text{AUC} = \int_{-\infty}^{\infty} \left(1 - F_P(u)\right) f_N(u)\,\mathrm{d}u = 1 - \int_{-\infty}^{\infty} F_P(u)\, f_N(u)\,\mathrm{d}u =$$

$$= 1 - \int_{-\infty}^{\infty} \left(\int_{-\infty}^{v} f_P(v)\,\mathrm{d}v\right) f_N(u)\,\mathrm{d}u = 1 - \Pr\left(X_P \leq X_N\right)$$

Métodos para calcular el AUC

- Método no paramétrico (regla trapezoidal)

$$\widehat{\mathrm{AUC}} = \sum_{t=1}^{T} \frac{1}{2} \left(\mathrm{FFP}_t - \mathrm{FFP}_{t-1}\right) \cdot \left(\mathrm{FVP}_t + \mathrm{FVP}_{t-1}\right)$$

- Método paramétrico (caso binormal)

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

Comparación de pruebas

- Contraste

$$H_0: \quad \mathrm{AUC}_1 = \mathrm{AUC}_2$$

$$H_1: \quad \mathrm{AUC}_1 \neq \mathrm{AUC}_2$$

- Estadístico

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

Aleatoriedad de la prueba

- Contraste

$$H_0: \quad \text{AUC} = 0{,}5$$

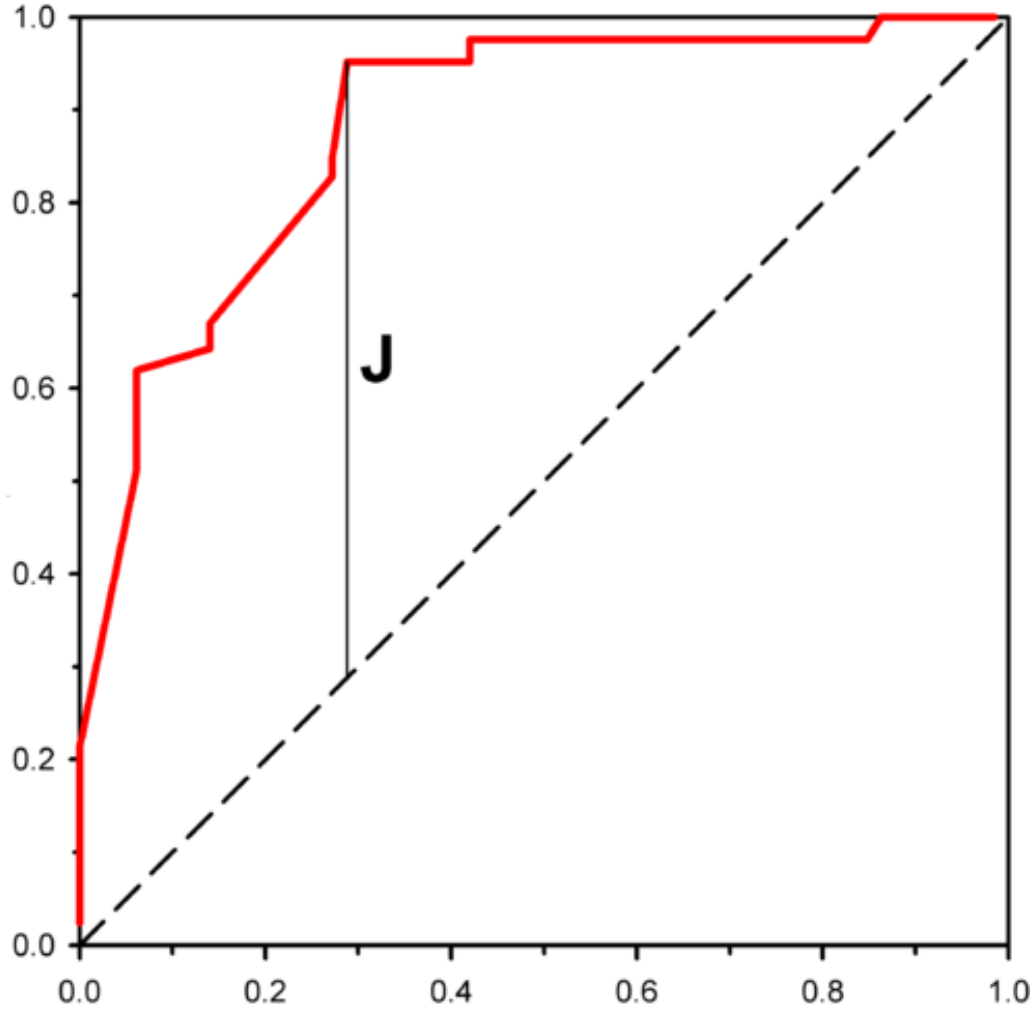
$$H_1: \quad \text{AUC} \neq 0{,}5$$

- Estadístico

$$z = \frac{\widehat{\text{AUC}} - 0{,}5}{\text{ET}\left(\widehat{\text{AUC}}\right)}$$

Punto de corte: método 1

- Usar el umbral correspondiente al índice de Youden



Punto de corte: método 2

- Buscar el punto $(1-E; S)$ sobre la curva ROC más cercano al punto $(0; 1)$
- La pendiente sobre ese punto vale

$$m = \frac{p \cdot \Pr(\text{falsos positivos})}{(1-p) \cdot \Pr(\text{falsos negativos})}$$

siendo p la prevalencia del evento en la población

Punto de corte: método 3

- En este caso se tienen en cuenta los costes de los dos tipos de error
- Buscar el punto $(1-E; S)$ sobre la curva ROC más cercano al que minimiza dichos costes
- La pendiente sobre ese punto vale

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

siendo p la prevalencia del evento en la población

Implementación ingenua

```
nN <- nP <- 50                                    #tamaños
N <- rnorm (nN, 0, 1)                             #negativos
P <- rnorm (nP, 1, 1)                             #positivos
x <- union (N, P)                                 #todos
umbral <- c (-Inf, x[-1]-diff(x)/2, Inf) #cortes posibles
s <- sapply (umbral,                              #sensibilidad
                function (c) {
                    vp <- sum (P >= c)
                    vp / nP })
```

Implementación ingenua

```
## 1 - especificidad

e <- 1 - sapply (umbral,
                 function (c) {
                   vn <- sum (N < c)
                   vn / nN })

o <- order (e, s)

plot (e[o], s[o], type="s")
```

Implementación mediante «ecdf»

```
nN <- nP <- 50                                          #tamaños
N <- rnorm (nN, 0, 1)                                   #negativos
P <- rnorm (nP, 1, 1)                                   #positivos
C <- 1 - ecdf (P) (sort (N, decreasing=TRUE)) #CORva
plot (C, type="s")
```

Biblioteca «pROC»

```
nN <- nP <- 50                                          #tamaños
N <- rnorm (nN, 0, 1)                                   #negativos
P <- rnorm (nP, 1, 1)                                   #positivos
install.packages ("pROC")
library ("pROC")
C <- roc (controls=N, cases=P)                          #CORva
plot (C)
coords(C,"best")    #mejor umbral según Youden (método 1)
coords(C,"best",best.method="closest.topleft") #método 2
## para método 3, opción «best.weights»
auc (C)
ci (C)                  #intervalo de confianza
C1 <- roc (rbinom(80,1,0.7), rnorm(80))
roc.test (C, C1)    #para comparar dos CORvas
```

Biblioteca «nsROC»

```
nN <- nP <- 50                          #tamaños
N <- rnorm (nN, 0, 1)                   #negativos
P <- rnorm (nP, 1, 1)                   #positivos
install.packages ("nsROC")
library (nsROC)
X <- c (N, P)                           #muestra completa
D <- c (rep(0,nN), rep(1,nP))           #0 = control, 1 = caso
G <- gROC (X, D, plot.roc=TRUE, plot.density=TRUE)
G$auc
G$roc
G$side                                  #unilateral por omisión
```

Biblioteca «nsROC»: dos umbrales

```
nN <- nP <- 50                          #tamaños
N <- rnorm (nN, 0, 1)                   #negativos
P <- rnorm (nP, 0, 5)                   #positivos
X <- c (N, P)                           #muestra completa
D <- c (rep(0,nN), rep(1,nP))           #0 = control, 1 = caso

G <- gROC (X, D, plot.roc=TRUE, plot.density=TRUE,
           side="both")
G                                       #explicación y AUC
```