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

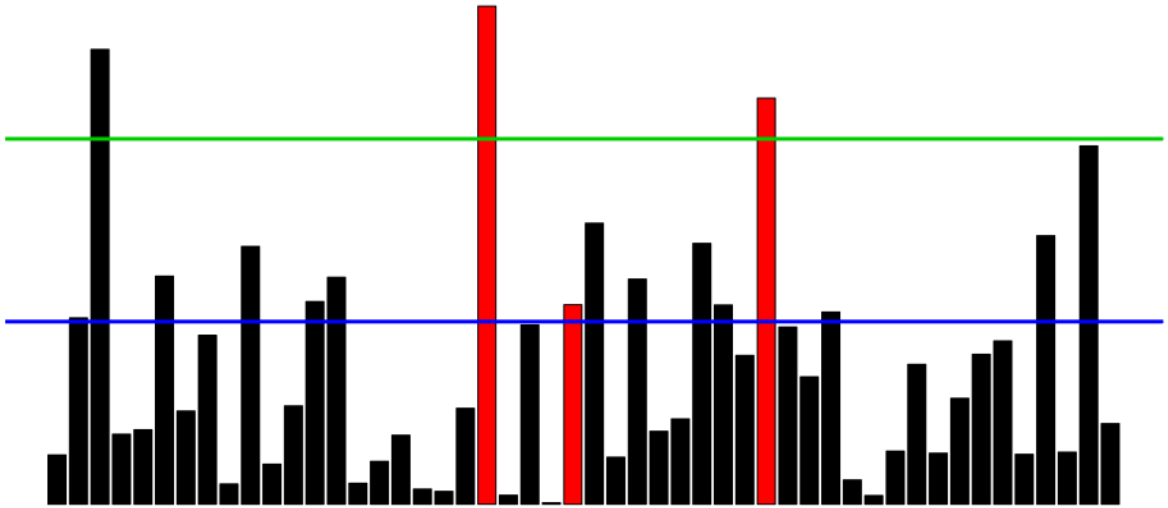
1. Teoría
   - Motivación, conceptos previos y definición
   - Inferencia
   - Elección del umbral
2. Práctica en R

Origen

- Gráfico estadístico
- Ayuda a definir un umbral de clasificación entre dos grupos
  - Positivos (los que verifican cierta condición)
  - Negativos (los que no la cumplen)
- C.O.R. = característica operativa del receptor
- Proviene de la teoría de detección de señales (discernir entre señal y ruido)

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)

Objetivo

- umbral de detección
  - umbral alto
    no se detectan objetos reales (Falsos Negativos)
  - umbral bajo
    muchos avisos falsos (Falsos Positivos)
- Es conveniente buscar un valor que equilibre ambas opciones.
- La curva ROC ayuda a buscar un umbral que haga poco frecuentes ambos tipos de clasificaciones erróneas.

Estado

- Grupo al que realmente pertenece cada individuo
- Variable aleatoria D que toma los valores

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

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

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

Variable medida

- Variable 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

- Toma dos valores posibles en función del valor de X

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

Estimador de la prevalencia

- Proporción muestral

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

- Depende del tipo de muestreo

Términos habituales

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

	Y=1	Y=0
D=1	VP	FN
D=0	FP	VN

Sensibilidad

- Probabilidad de acertar en la clasificación de un individuo positivo

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

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

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

Especificidad

- Probabilidad de acertar en la clasificación de un individuo negativo

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

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

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

Fracciones (tabla resumen)

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

Valores predictivos

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

Definición de curva ROC

- Curva ROC **poblacional**
- Gráfica que representa, para cada posible valor del umbral c, la Sensibilidad frente a $1 -$ Especificidad, es decir,
  - Abscisas $= 1 -$ Especificidad
  - Ordenadas $=$ Sensibilidad
- Suponiendo que X toma, en general, valores mayores para los positivos,

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

donde

- $t = 1 - E(x) = \Pr\left(Y(x) = 1 \mid D = 0\right) = 1 - F_N(x) =$ $\Pr\left(\text{Error} \mid \text{Negativo } D = 0\right)$
- $ROC(t) = S(x) = 1 - F_P(x) = \Pr\left(\text{Acierto} \mid \text{Positivo } D = 1\right)$

Método no paramétrico para estimar la curva ROC

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

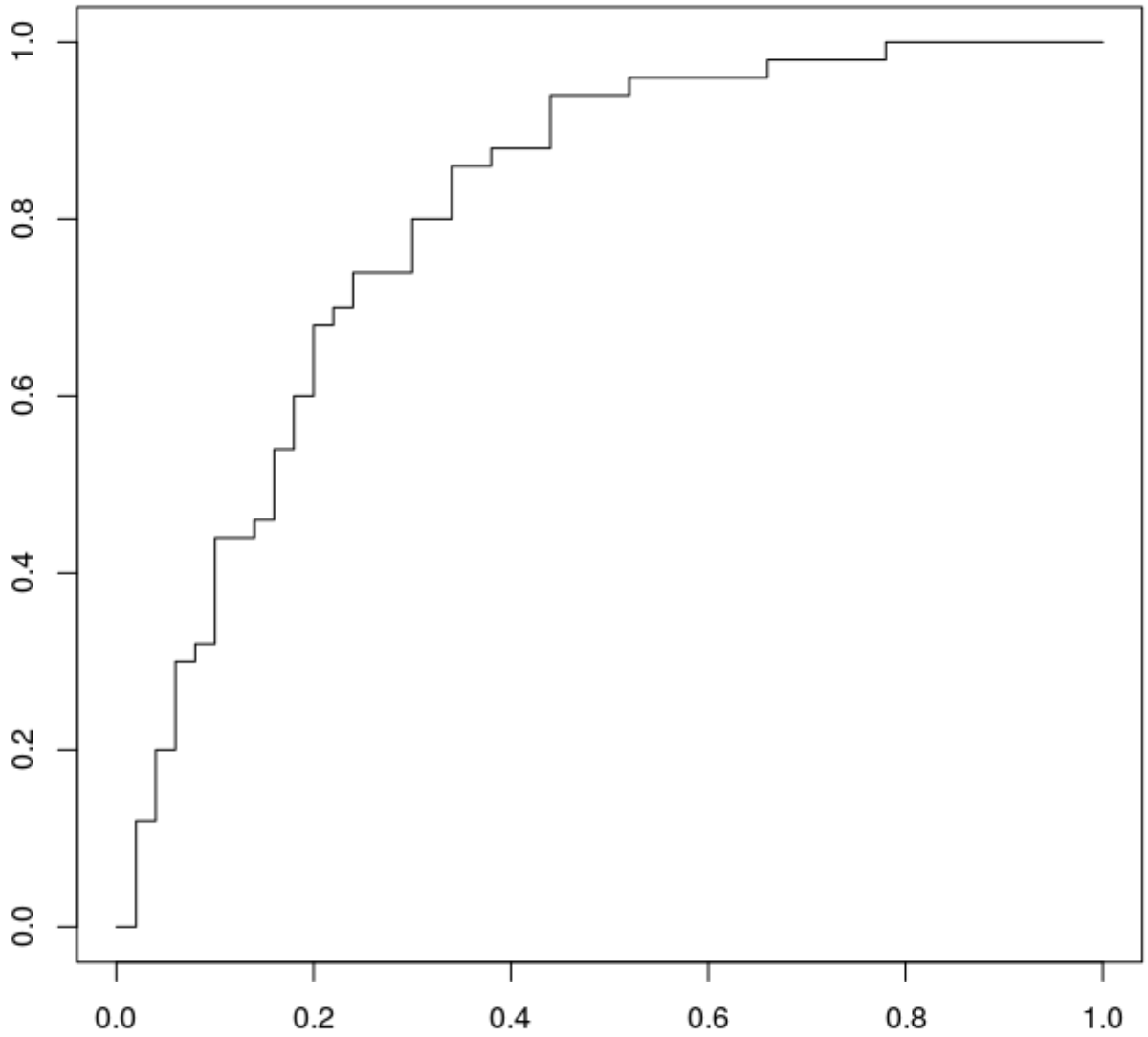
$$\hat{F}_P(x) = \frac{\text{nº}(x_i \leq x \mid Positivos)}{n_P}$$
$$\hat{F}_N(x) = \frac{\text{nº}(x_i \leq x \mid Negativos)}{n_N}$$

- Se define la curva ROC empírica como

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

- 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

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

donde $\alpha = \dfrac{\mu_N - \mu_P}{\sigma_P}$, $\beta = \dfrac{\sigma_N}{\sigma_P}$,
Φ = función de distribución de la gausiana típica

Método paramétrico para estimar la curva ROC

- La curva ROC estimada es

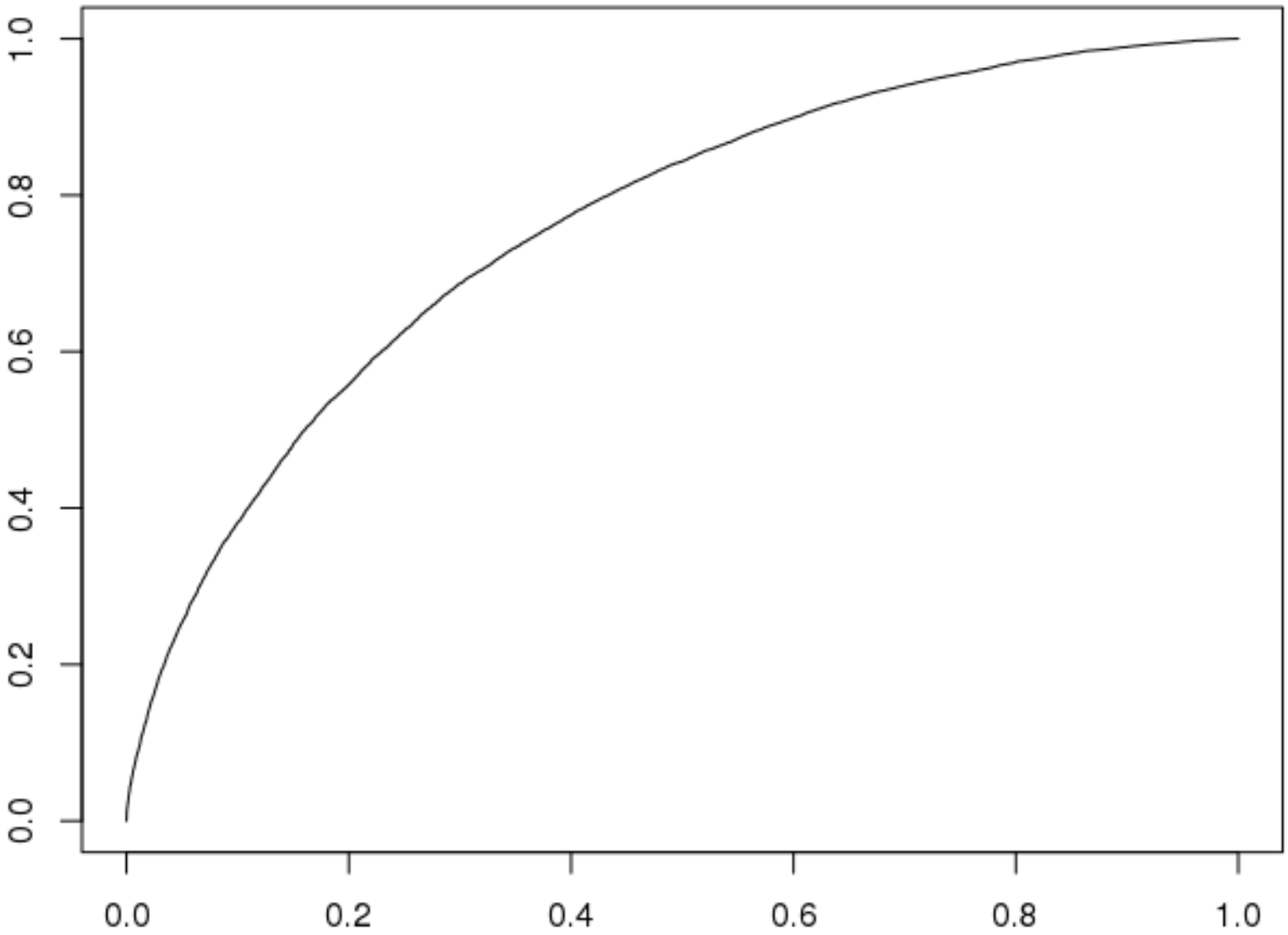
$$\widehat{ROC}(t) = 1 - \Phi\left(\hat{\alpha} + \hat{\beta} \cdot \Phi^{-1}(1-t)\right) \qquad 0 \leq t \leq 1$$

donde

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

- Tiene un aspecto liso

Método no 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:

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

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 $\widehat{\Pr}(\textit{acertar}) = \dfrac{VP + VN}{n} = \dfrac{\text{resultados acertados}}{\text{total de la muestra}}$
- Proporción de aciertos en la clasificación sin distinguir positivos de negativos

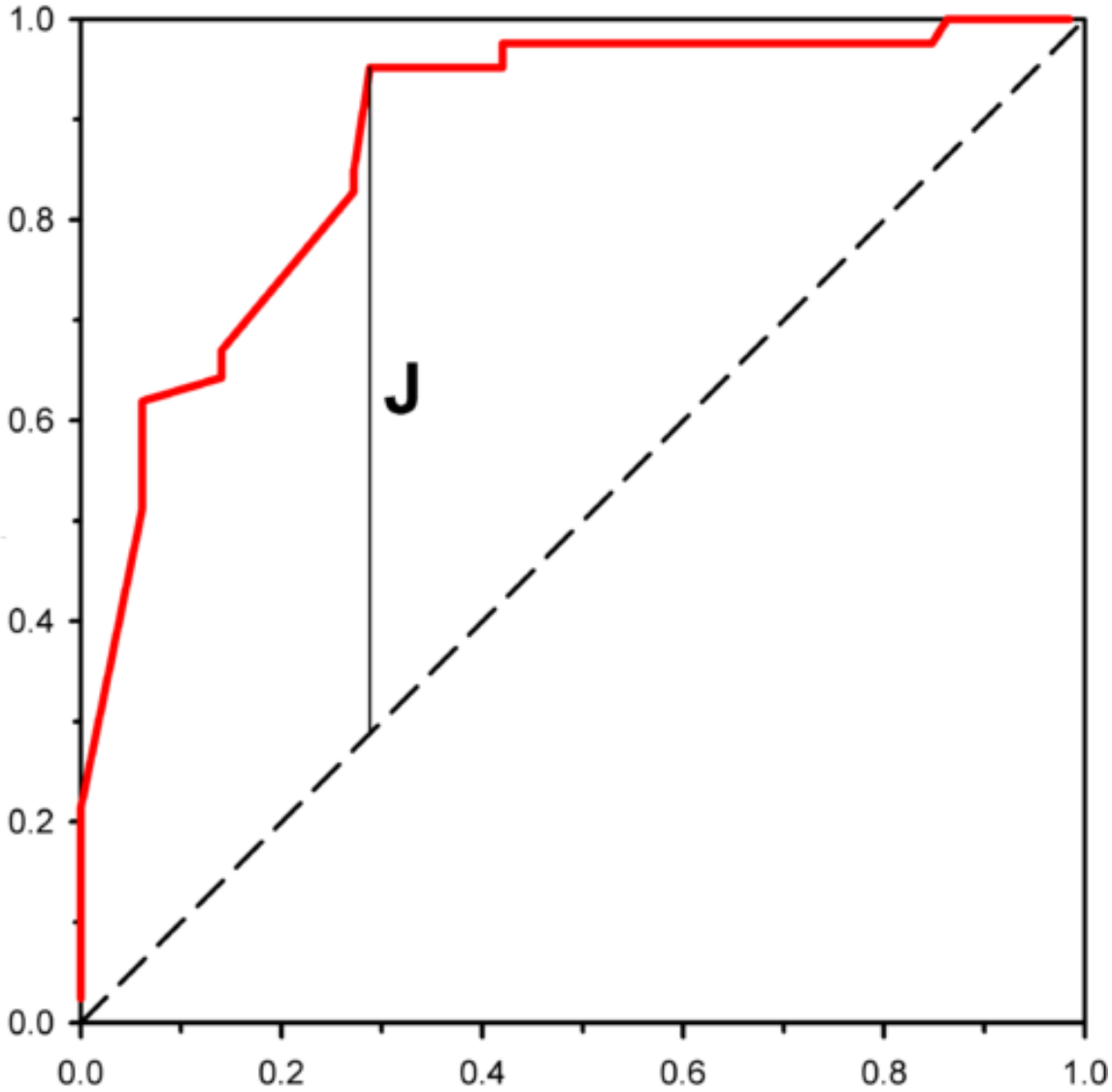
Índice de Youden

- 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 $\hat{J} = FVP - FFP$

Índice de Youden



Área bajo la curva

- AUC (*area under curve*) se define así:

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

- $0{,}5 \leq AUC \leq 1$
- $AUC \approx 0{,}5 \iff$ 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)

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

- Método paramétrico (caso binormal)

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

Comparación de pruebas

- Contraste

$$H_0: \quad AUC_1 = AUC_2$$

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

- Estadístico

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

Aleatoriedad de la prueba

- Contraste

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

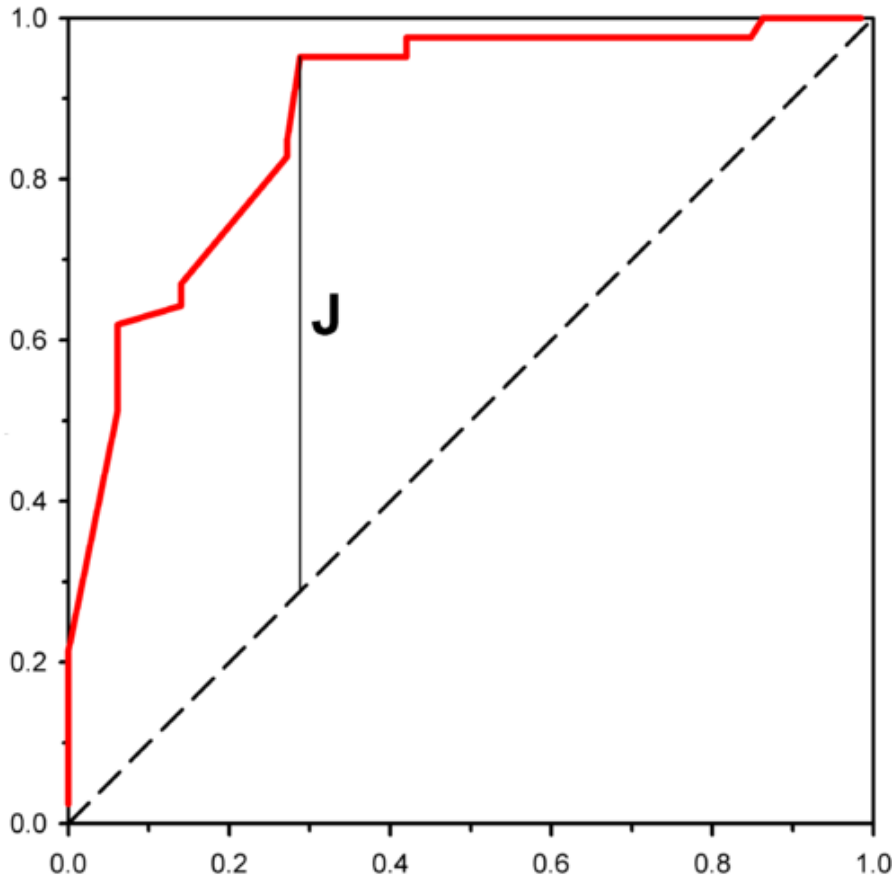
$$H_1: \quad AUC \neq 0{,}5$$

- Estadístico

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

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
t <- sort (union (N, P))                # todos
U <- c (-Inf, t[-1]-diff(t)/2, Inf)     # umbrales
s <- sapply (U,                         # sensibilidad
             function (Ui)
               sum (P > Ui) / nP)
e  <- 1 - sapply (U,                    # 1 - especificidad
                  function (Ui)
                    sum (N <= Ui) / nN)
o <- order (e, s)                       # 's' rompe empates
plot (e[o], s[o], type="s")             # "s" para peldaños
```

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 (seq (0, 1, length.out=nN), 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)
auc (C)                 # área bajo la curva
ci (C)                  # intervalo de confianza
roc.test (C1, C2)       # para comparar dos CORvas
```

Biblioteca «pROC»: mejor umbral

```
plot (C)

## mejor umbral según método 1 (Youden)
uY <- coords (C, "best")
uY <- coords (C, "best", best.method = "youden") # ídem
abline (v = uY["specificity"], col = 2)          # rojo

## mejor umbral según método 2 (cercano a la esquina)
uC <- coords (C, "best", best.method="closest.topleft")
abline (v = uC["specificity"], col = 3)          # verde

## para método 3, la opción «best.weights»
## permite dar los dos coeficientes de costes
```

Biblioteca «nsROC»

```
nN <- nP <- 50                          # tamaños
N <- rnorm (nN, 0, 1)                   # negativos (controles)
P <- rnorm (nP, 1, 1)                   # positivos (casos)

install.packages ("nSROC")              # © Sonia Pérez
library (nsROC)                         # Dpt° Estadística, UO

X <- c (N, P)                           # marcador
D <- c (rep (0, nN), rep(1, nP))        # 0=control  1=caso

gROC (X, D, plot.roc = TRUE, plot.density = TRUE)
```