R による ROC 分析法

公衆衛生学教室セミナー　2007 年 1 月 11 日　中澤　港

1　ROC 分析とは

ROC 曲線とは，Receiver Operating Characteristic 曲線の略である[*1]。集団を対象に，すばやく実施可能な方法で，疾病を暫定的に識別することをスクリーニングというが，いくつかのスクリーニング方法があるときに，それらの相対的な有効性を視覚的に判定する基準の一つが ROC 曲線である。

１つのスクリーニング方法について陽性・陰性の基準値を最小値から最大値まで段階的に変えると，偽陽性率（= 1 − 特異度）も感度（病気の人を正しく陽性と判定する割合）も 0 から 1 まで変わるので，偽陽性率を横軸に，感度を縦軸にとって線で結ぶと，基準値の変化に対応する曲線を引くことができる。この曲線ができるだけ左上を通る方がスクリーニングとしての有効性は高い方法だといえる。

また，この曲線の最も左上の点（理想は偽陽性率 0 で感度 1 だが，現実にそうなることはまずない）を与える基準値が，陽性・陰性を分けるカットオフポイントとして最も有効性が高いと判断される。

つまり，ROC 曲線は，ある検査値について適切なカットオフポイントを検索するのにも使えるし，複数のスクリーニング方法の優劣を比較することにも使える。

ROC 曲線を描いて視覚的評価をするだけではなく，AUC（Area under curve; 曲線下面積）を計算する，あるいは右下の点からもっとも離れた点を与えるカットオフポイントを最適値とするなどの計算も含めて，ROC 分析と呼ぶ。複数のスクリーニング方法の AUC を比較し，最も大きい AUC を与える方法が最も優れていると考えるのが普通である。ただし，感度や特異度が最も優れていても，他のもっと廉価に大勢を検査できる方法と大差なければ，高価だったり時間や手間がかかる（倫理面も含めて）などの理由で採用されない場合もある。

2　Excel でやってみる

具体例で考えよう。以下のデータ（架空である）が得られたとする。

対象者	質問紙得点	臨床診断
1	20	うつ
5	22	うつ
6	28	うつ
2	13	健康
3	19	健康
4	21	健康
7	11	健康
8	25	健康
9	16	健康
10	19	健康

[*1] 日本語では，受診者動作特性曲線という訳語がついている教科書と，受信者動作特性曲線という訳語がついている教科書が並立しているが，ROC が何の略であるかを明示して「ROC 曲線」だけを掲載している本も増えてきたので，ここでも敢えて訳さないことにする。手元にある本で調べると，日本疫学会（編）「疫学　基礎から学ぶために」南江堂，能登洋「日常診療にすぐに使える臨床統計学」羊土社などが「受診者」派で，鈴木・久道（編）「シンプル衛生公衆衛生学 2006」南江堂，日本疫学会（訳）「疫学辞典　第３版」日本公衆衛生協会，フレッチャー RH，フレッチャー SW，ワグナー EH，福井次矢（監訳）「臨床疫学」メディカルサイエンスインターナショナルなどが「受信者」派であった。稲葉・野崎（編）「新簡明衛生公衆衛生　改訂４版」南山堂，丹後俊郎「メタ・アナリシス入門」朝倉書店などは，「ROC 曲線」だけを掲載していた。

この質問紙得点が，あるカットオフポイントより高いことを，うつのスクリーニングとして使おうというのが，このデータを得た目的であるとすると，問題は，適切なカットオフポイントを見つけることになる。

例えば，カットオフポイントを 18，すなわち，質問紙得点が 18 点以上なら陽性，そうでないなら陰性と判定することにすると，以下のクロス集計表ができる。

	うつ	健康
陽性	3	4
陰性	0	3

このとき，感度は $3/(3+0)=1$，特異度は $3/(4+3)=0.429$，偽陽性率は $4/(4+3)=1-0.429=0.571$ となる。得点の最小値から最大値+1 までカットオフポイントをずらしていくと，感度も偽陽性率も 1 から 0 まで変化するので，これをグラフに描けば ROC 曲線となる[*2]。ただ，Excel では AUC を計算したり最適カットオフポイントを見つけることは容易ではないし，いちいち多くのセルを使って計算式を入力するのも面倒である。

3 自分で関数を作る方法

そこで R の登場となる。ROC 曲線の変曲点はデータがあるところであることを考慮し，素直に式を書けば，次の枠内のような関数 roc を定義することができる。カットオフポイントをずらしていったときの，感度，偽陽性率，感度 0 偽陽性率 1 の点からの距離をリストとして返す。

```
roc <- function(values,iscase) {
 cutoffs <- unique(sort(values))
 cutoffs <- c(cutoffs,max(values)+1)
 ns <- length(cutoffs)
 sensitivity <- rep(0,ns)
 falsepositive <- rep(0,ns)
 dist <- rep(0,ns)
 for (i in 1:ns) {
  cutoff <- cutoffs[i]
  positives <- ifelse(values >= cutoff,1,0)
  D <- sum(iscase==1)
  H <- sum(iscase==0)
  PinD <- sum(positives==1 & iscase==1)
  NinH <- sum(positives==0 & iscase==0)
  sensitivity[i] <- PinD/D
  falsepositive[i] <- 1-NinH/H
  dist[i] <- sqrt((PinD/D)^2+(NinH/H)^2)
 }
 list(cutoffs,sensitivity,falsepositive,dist)
}
```

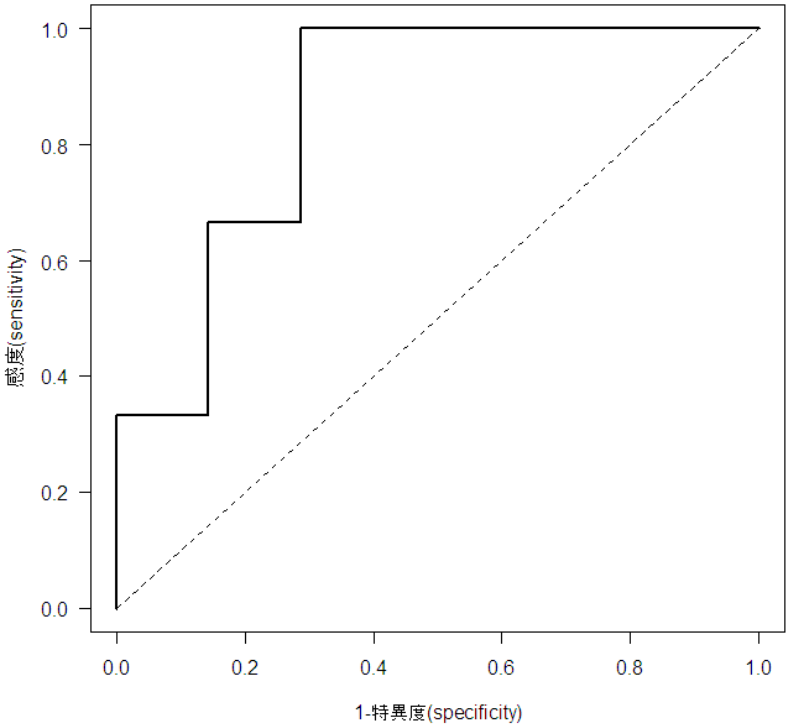
結果として得たリストの値を使って ROC 曲線を描き，最適カットオフポイントを求めるには次の枠内のように関数定義する。

[*2] 丹後俊郎『メタ・アナリシス入門』朝倉書店に紹介されているように，クロス集計表のどこかのセルが 0 になる場合は各セルに 0.5 ずつ加える Woolf(1955) の修正を薦める教科書もあるが，その場合曲線の端点が (0,0) と (1,1) にならないので，本稿では修正しない。

```
rocc <- function(...) {
 res <- roc(x,y)
 cat("cutoff\tsensitivity\t1-specificity\tdistance\n",
 sprintf("%5.3f\t%5.3f\t%5.3f\t%5.3f\n",res[[1]],res[[2]],res[[3]],res[[4]]))
 mlcs <- "最適カットオフポイント:%5.3f，感度:%5.3f，特異度%5.3f\n"
 mlcc <- which.max(res[[4]])
 cat(sprintf(mlcs,res[[1]][mlcc],res[[2]][mlcc],1-res[[3]][mlcc]))
 plot(res[[3]],res[[2]],type="l",lwd=2,xlab="1-特異度 (specificity)",ylab="感度 (sensitivity)")
 lines(c(0,1),c(0,1),lwd=1,lty=2)
}
```

最初に示したデータを使って計算するには次の枠内を打つ。最初の 2 行はデータ入力だが，Excel 上でセルを選んでコピーし，`x <- scan("clipboard")` のようにしても入力可能である。このように関数定義をすれば，実行は `rocc(x,y)` だけで済む。最適カットオフポイント 20，そのときの感度が 1 で特異度が 0.714 とわかる。

```
x <- c(20,22,28,13,19,21,11,25,16,19)
y <- c(rep(1,3),rep(0,7))
rocc(x,y)
```



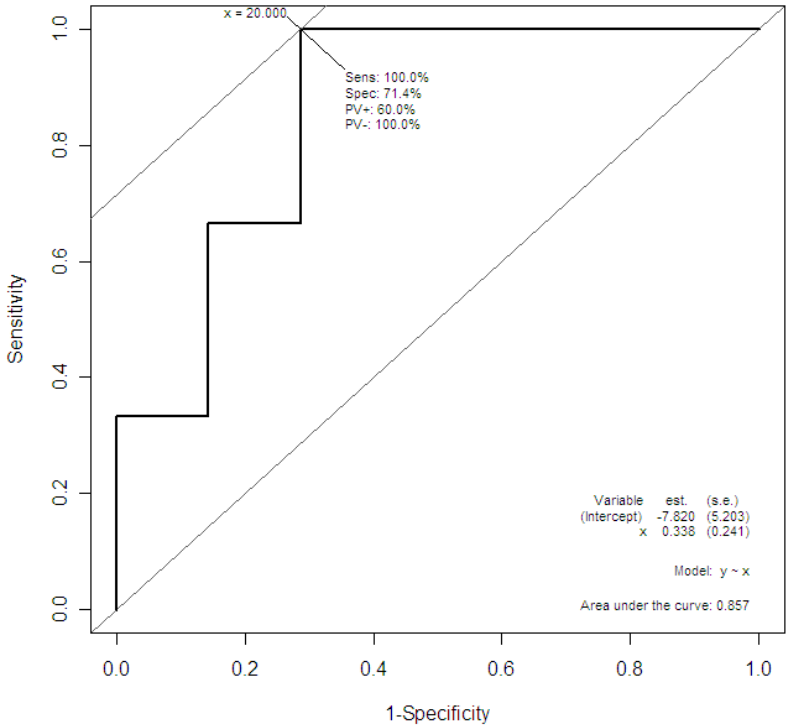
上記の方法は中身を完全に把握できるところはいいが，今ひとつ美しくないし，AUC の計算をさぼっている。そこで，CRAN から 1 つ，Epi というライブラリをインストールすることにする。インストール方法は，群馬大学内でインターネットに接続されたコンピュータであれば R のコンソールで `install.packages("Epi",dep=T)` とすればよい。もしレポジトリあるいはミラーを選ぶようにという選択肢がでてきたら，Japan(Tsukuba) または Japan(Tokyo) を選べばよい。

4　Epi ライブラリを使う方法

Epi ライブラリは，デンマーク・コペンハーゲン大学の Bendix Carstensen らが開発して CRAN で公開している，慢性疾患の疫学のためのライブラリである[*3]。ROC の他，age-period-cohort モデルや Lexis diagram を描く関数も含まれている。

Epi ライブラリを使った実行方法は非常に簡単で，次の枠内を打つだけでいい。結果も図内にすべて示される。

```
require(Epi)
ROC(x,y,plot="ROC")
```



[*3] 詳しくは http://staff.pubhealth.ku.dk/~bxc/Epi/を参照。