

RCD曲線、ROC曲線

動脈硬化症を例に説明します。

動脈硬化は、血圧、コレステロール値、中性脂肪値、脈波、血糖値、尿酸値などを計測し、総合的に判断するものとされています。

下表の左側は動脈硬化症群と判定された人、右側は正常と判定された人の総コレステロールを記述しています。

一般に、総コレステロールが 220mg/dl 以上で高コレステロール血症と判定されることが多いようです。

動脈硬化症群	総コレステロール(mg/dl)
動硬症-1	223
動硬症-2	227
動硬症-3	231
動硬症-4	233
動硬症-5	237
動硬症-6	239
動硬症-7	240
動硬症-8	243
動硬症-9	249
動硬症-10	251

正常群	総コレステロール(mg/dl)
正常-1	201
正常-2	203
正常-3	208
正常-4	211
正常-5	213
正常-6	215
正常-7	223
正常-8	228
正常-9	229
正常-10	230

2つの表を見ると、動脈硬化症者は全員 220mg/dl 以上ですが、正常者の中にも 220 mg/dl 以上が 4 人います。

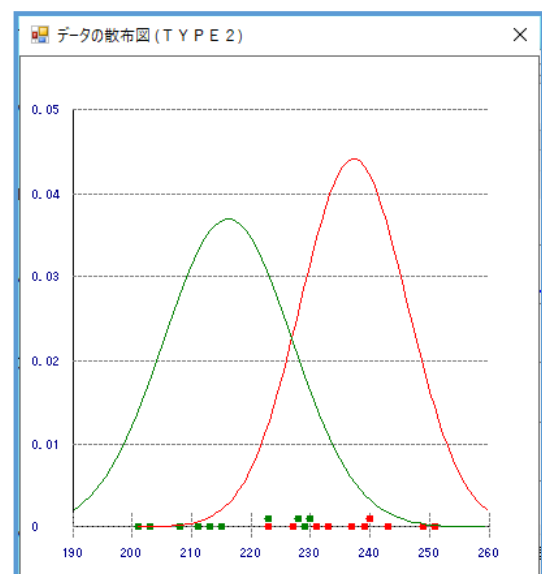
右のグラフは、横軸をコレステロール値として個々のデータをプロットしたものです。
(縦軸はそれぞれを正規分布の標本として描画した確率分布の確率値)

グラフの

- ・赤の点は左表の動脈硬化症患者
- ・緑の点は右表の正常者

を表しています。

コレステロール値が 220～230 あたりには、両方が混じっているために、しきい値を特定して、それを境に陽性か陰性かを確実に峻別することはできません。



これはコレステロール値だけで、動脈硬化を判定することの難しさを示します。

このように、本来はいくつかの指標を元に判定すべきところで、その場合は、別途解説する **判別分析** で 評価をしていくことになります。

しかし、1つの指標のみしか与えられていない、そしてそこから判断が求められるという状況において、あえて指標のしきい値に何を指定すべきかを考えます。

データがしきい値以上を 陽性、しきい値未満を 陰性 とします。
もちろん逆にしきい値以下を陽性とみなす合もありますが、ここでは、しきい値以上を陽性 として 話を進めます。

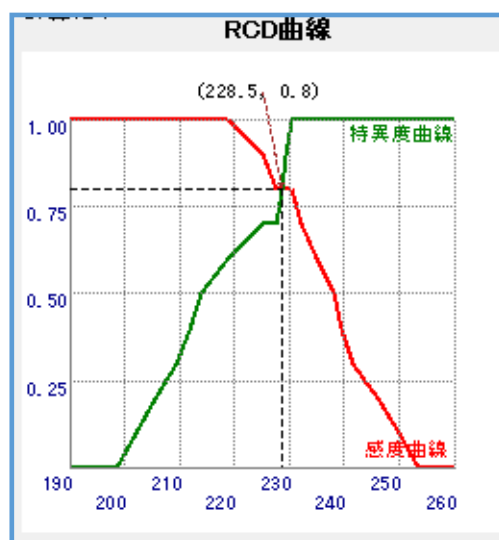
あるしきい値を設定したら、陽性、陰性の判断が 100%の確度で決定できればよいのですが、先の例では それはできません。
しきい値を 220mg/dl とした場合、動脈硬化症群（左表）は全員正しく陽性と判定されますが、正常群（右表）の 4 人は陽性と判定され、偽陽性が 4 人となります。
今度はしきい値を 231mg/dl とした場合、正常群（右表）は全員正しく陰性と判定されますが、動脈硬化症群（左表）の 2 人は陰性と判定され、偽陰性が 2 人となります。
「動脈硬化症群」（左表）と正常群（右表）とで、データが被っている以上、仕方のないことです。

ここで二つの言葉を定義します。

感度 : 真の陽性者が正しく陽性と判断される割合。0.0~1.0 の値を取ります。

特異度 : 真の陰性者が正しく陰性と判断される割合。0.0~1.0 の値を取ります。

しきい値を最も小さい値 201 mg/dl とすれば、感度は 1.0 で、特異度は 0.0 です。
逆に最も大きい値 251 mg/dl とすれば、感度は 0.0 で、特異度は 1.0 です。
そこで、しきい値を変数として、小さい値から大きい値に変化させて、感度と特異度がどのように変化するかをグラフで表現したものが以下です。



図を見てわかるように、感度（赤で表示）はしきい値が小から大に変化するにつれ、1.0 から 0.0 に減少します。逆に特異度（緑で表示）は 0.0 から 1.0 に増加します。
この図を **RCD曲線**（Relative Cumulative frequency Distribution）といいます。

この図の2つの曲線の交点 228.5 をしきい値とすると、感度、特異度ともに 0.8 となり、ある意味で ” 妥当な ” 結果を得ることになります。

R C D曲線 を利用して、R O C曲線 (Receiver Operating Characteristic) を作成します。

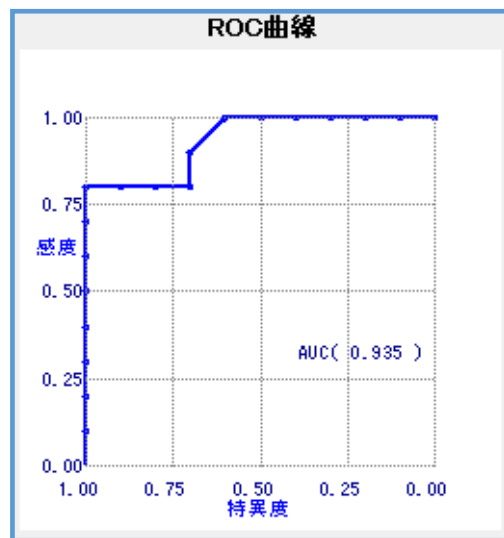
R O C曲線は横軸に特異度を取ります。ただし値は左から右へ、1.0 から 0.0 へと変化します。

ですので横軸を非特異度という場合もあります。(非特異度 = 1 - 特異度)
縦軸は感度を取ります。値は下から上へ、0.0 から 1.0 へと変化します。

R C D曲線の、右側から左側に向かって以下のように、逐次データを採取します。

しきい値=260	:	(感度, 特異度) = (0.0 , 1.0)
しきい値=250	:	(感度, 特異度) = (0.11, 1.0)
しきい値=240	:	(感度, 特異度) = (0.40, 1.0)
しきい値=230	:	(感度, 特異度) = (0.80, 0.89)
.		
しきい値=190	:	(感度, 特異度) = (1.0 , 0.0)

上記のデータをR O C曲線図の感度と特異度としてプロットしたものが以下です。



R O C曲線は、陽性、陰性の判定の指標として用いているコレステロール値が、判定に役立つものかを視覚的に表現します。

R O C曲線が 上図のように、左上に山をなしている場合は、判定として役立っていることを示します。

また、AUCが0.935 と表示されていますが、これも 判定として役立っているかを示す指標で 0.0 ~1.0 の値を取り、1.0 に近いほど、“判定に役立っている”と判断されます。

AUCとはArea Under Curve の略で、R O C曲線の下側部分の面積です。

今回の例では、コレステロール値が 220~230 のあたりで、陽性と陰性とでデータがかぶっているために、AUCは1.0 にはなりませんでした。

陽性と、陰性との間にデータの重複がない場合、AUCはどうなるでしょうか？

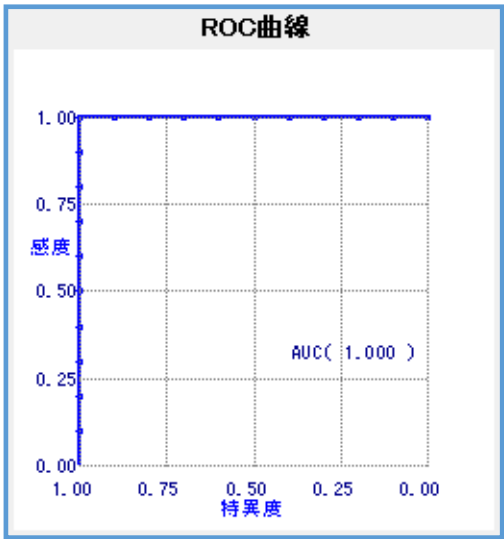
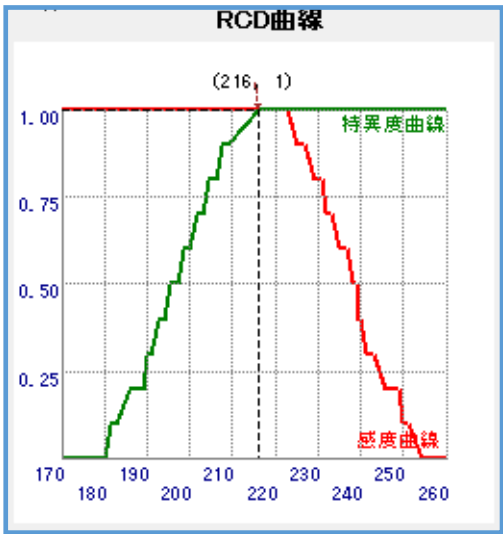
以下のデータで考えます。

動脈硬化症群	総コレステロール(mg/dl)
動硬症-1	223
動硬症-2	227
動硬症-3	231
動硬症-4	233
動硬症-5	237
動硬症-6	239
動硬症-7	240
動硬症-8	243
動硬症-9	249
動硬症-10	251

正常群	総コレステロール(mg/dl)
正常-1	180
正常-2	183
正常-3	189
正常-4	191
正常-5	194
正常-6	197
正常-7	200
正常-8	203
正常-9	206
正常-10	209

この場合、しきい値を 209 超かつ 223 以下に指定すれば、感度、特異度共に 1.0 です。

そして AUC は 1.0 です。



陽性グループと、陰性グループが 同一データの場合はどうなるでしょう？

陽性グループ

動脈硬化症群	総コレステロール(mg/dl)
動硬症-1	223
動硬症-2	227
動硬症-3	231
動硬症-4	233
動硬症-5	237
動硬症-6	239
動硬症-7	240
動硬症-8	243
動硬症-9	249
動硬症-10	251

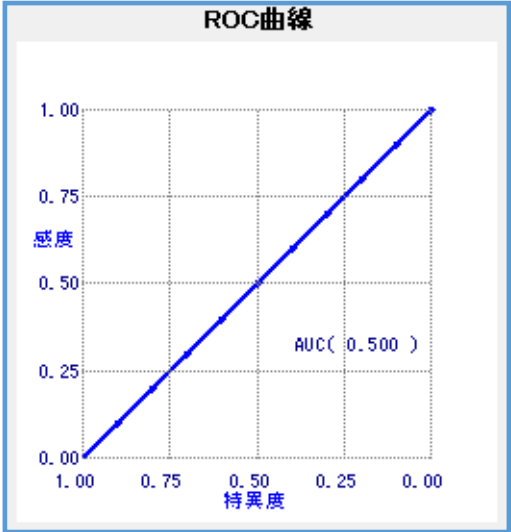
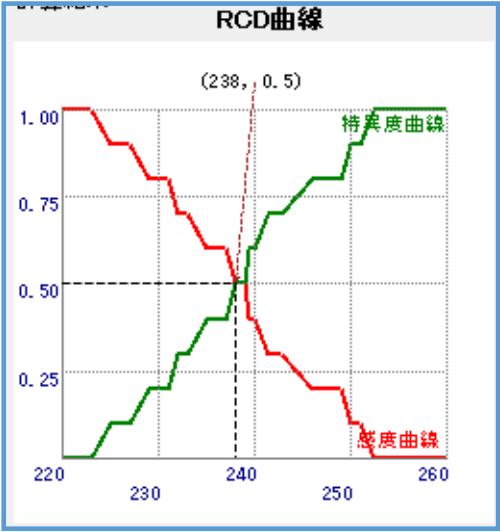
陰性グループ

動脈硬化症群	総コレステロール(mg/dl)
動硬症-1	223
動硬症-2	227
動硬症-3	231
動硬症-4	233
動硬症-5	237
動硬症-6	239
動硬症-7	240
動硬症-8	243
動硬症-9	249
動硬症-10	251

この場合、しきい値を動かしても、感度、特異度が共に妥当な値は見つかりません。

ROC 曲線は 直線 となります。

AUC は 0.5 です。



この場合は、この判定が役に立たない ことを示しています。

陽性グループと陰性グループとを逆に指定してしまった場合はどうなるでしょう？

陽性グループ

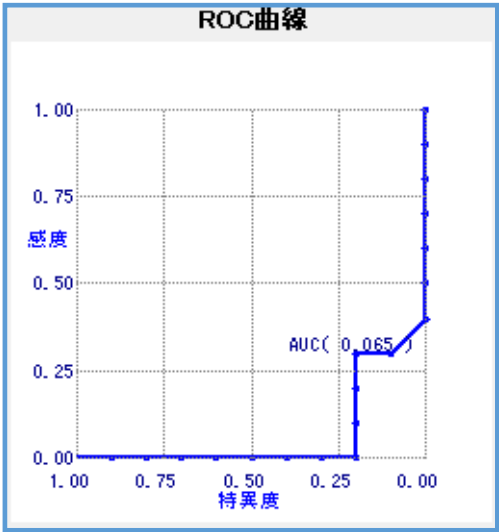
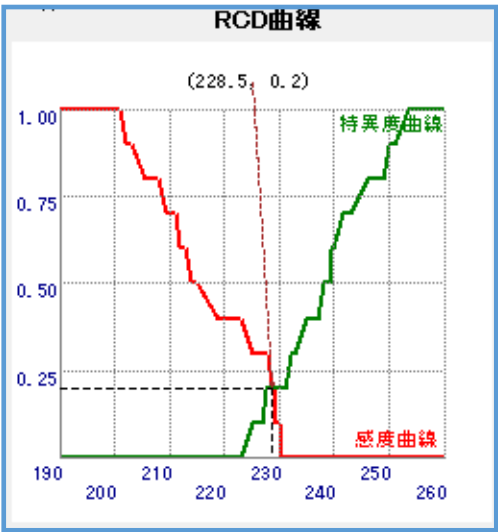
正常群	総コレステロール (mg/dl)
正常-1	180
正常-2	183
正常-3	189
正常-4	191
正常-5	194
正常-6	197
正常-7	200
正常-8	203
正常-9	206
正常-10	209

陰性グループ

動脈硬化症群	総コレステロール (mg/dl)
動硬症-1	223
動硬症-2	227
動硬症-3	231
動硬症-4	233
動硬症-5	237
動硬症-6	239
動硬症-7	240
動硬症-8	243
動硬症-9	249
動硬症-10	251

この場合は、ROC 曲線は 右下向きの曲線 となります。

AUC は 0.065 です。



この場合は、この判定が意味をなさない ことを示しています。

データの扱いを 誤ると こういうことになります。