Analyse von ROC-Kurven mit Hilfe von SAS

Hans-Peter Altenburg
Siemens Healthcare Diagnostics
Products GmbH, QA/SV
Emil-von-Behring-Str. 76
35001 Marburg
hans-peter.altenburg@siemens.com

Zusammenfassung

Die Verwendung von ROC-Kurven („receiver operating characteristic"-Kurven) sind ein statistisches Werkzeug um in diagnostischen Verfahren die Genauigkeit von Vorhersagen zu bestimmen. Sie werden nicht nur in der medizinischen Diagnostik sondern auch in modernen Data-Mining Fragestellungen wie Credit-Scoring oder auch Wettervorhersagen verwendet. Alle diese Anwendungen haben eines gemeinsam: Vorhersagen werden auf der Basis von Prädiktoren gemacht, bevor der eigentliche „vorherzusagende" Wert beobachtet wird. Es wird gezeigt, welche Möglichkeiten SAS bietet, um solche Fragestellungen zu bearbeiten. Direkt auf die Fragestellungen zugeschnittene Lösungen gibt es in SAS nur für einfache Fragen. Für die meisten Anwendungen müssen vielmehr existierende Prozeduren auf die Erfordernisse zugeschnitten werden. Wie dies geschehen kann, soll im Beitrag in konkreten Beispielen aufgezeigt werden. Grundlegende Verfahren (wie diagnostische Kennzahlen) werden mit Hilfe der SAS-Prozeduren FREQ und LOGISTIC beschrieben. Für die Darstellung von ROC-Kurven stellt SAS Institute ein Makro zur Verfügung. Für Vergleich und Kovariablen-Adjustierung über binormale Modelle oder Regressionsmodelle können die Prozeduren NLMIXED, für Bootstrapping-Verfahren zur Erstellung von Konfidenzintervallen die Prozedur SURVEYSELECT verwendet werden.

Schlüsselwörter: Sensitivität, Spezifität, Receiver Operating Characeristic, AUC, ROC-Kurven-Vergleich, Binormal-Modell, Bootstrap, FREQ-Prozedur, LOGISTIC-Prozedur, NLMIXED-Prozedur, SURVEYSELECT-Prozedur

1 Einleitung

ROC Kurven (Receiver Operating Characteristic Kurven) sind ein wichtiges Werkzeug in der quantitativen Diagnostik. Eine andere (nicht so gebräuchliche) Bezeichnung lautet *Relative Operating Characteristic* Kurve. Die Vorgehensweise wurde entwickelt in den 1950-er Jahren als Nebenprodukt von Untersuchungen mit Radiosignalen, welche durch sog. „weisses Rauschen" gestört wurden.
Die Prozedur für einen neuen quantitativen Test verläuft normalerweise folgendermaßen ab: Führe einen neuen Test (mit einem evtl. neuen Produkt) anhand eines vorgegebenen Cutpoints für den quantitativen Test durch und erfasse die Ergebnisse. Verwende einen „Goldstandard" (z.B. Erkrankung liegt tatsächlich vor), um zu entscheiden, ob wirklich eine „Erkrankung" vorliegt und erfasse auch hier die Ergebnisse. Fasse die Ergebnisse beider Untersuchungen in einer Tabelle wie folgt zusammen.

Tabelle 1.1: Zusammenfassung der Stichproben- / Testergebnisse

		nicht erkrankt	***erkrankt***	**Σ**
		D +	D -	
> Cutpoint	T +	**a**	**b**	a+b
≤ Cutpoint	T -	**c**	**d**	c+d
Σ		a+c	b+d	N=a+b+c+d

Grundlegende diagnostische Kennzahlen sind dann:

Sensitivität = a/(a+c), Falsch Negative = c/(a+c),
Spezifität = d/(b+d), Falsch Positive = b/(b+d).

In einer „endlichen“ Population lassen sich diese relativen Häufigkeiten als bedingte Wahrscheinlichkeiten interpretieren und quasi in eine „Diagnostische Sprache” übersetzen, deren Begriffe wir im Folgenden dieses Artikels übernehmen wollen.

Tabelle 1.2: Bedingte Wahrscheinlichkeiten und Begriffe

Bedingte Wahrscheinlichkeit	***Bezeichnung***	***„englische Bezeichnung”***
P(T+ | D+)	Sensitivität (Se)	TPF (true positive fraction)
P(T- | D-)	Spezifität (Sp)	TNF (true negative fraction)
P(T+ | D-)	Falschpositive (F+)	FPF (false positive fraction)
P(T- | D+)	Falschnegative (F-)	FNF (false negative fraction

Anstelle der absoluten Häufigkeiten in Tabelle 1.1 könnten auch die entsprechenden diagnostischen Kennzahlen (Se, F+, F-, Sp) zusammengefasst angegeben werden (allerdings ohne die Summenwerte). In vielen Anwendungen werden die Kennzahlen auch in eine umgangssprachlich besser einprägsame Form gebracht. Hier ein paar Beispiele:

Tabelle 1.3: Umschreibung diagnostischer Kennzahlen

Diagnostische Kennzahl	***Umschreibung***
Sensitivität	Anteil erkannter Kranker, Anteil richtig Positiver
Spezifität	Anteil erkannter **Nicht**kranker, Anteil richtig Negativer
Falschnegative	Anteil übersehener Kranker
Falschpositive	Anteil falsch verdächtigter **Nicht**kranker

2 Bestimmung einfacher Kennzahlen mit der Prozedur FREQ

Das folgende einfache Zahlenbeispiel in Tabelle 2.1 soll die Vorgehensweise mit SAS erläutern.

Tabelle 2.1: Zahlenbeispiel

		erkrankt	***nicht erkrankt***
		+	-
Test	+	11	4
	-	2	6

Die Daten werden zeilenweise eingelesen. Über ein Format werden die Beschriftungen umgesetzt:

```
PROC FORMAT ;
     VALUE YesNo 0=' - '
                 1=' + ' ;
RUN ;
DATA study1 ;
     INPUT Test  Illness Count ;
FORMAT Test Illness YesNo.
  ;
DATALINES ;
     1 1 11
     1 0  4
     0 1  2
     0 0  6
     ;
PROC PRINT DATA=study1    ;
     ID  Test                  ;
     VAR  illness  Count       ;
RUN ;
```

Die Bestimmung der Kennzahl ***Sensitivität*** erfolgt dann über die SAS-Prozedur FREQ

```
PROC FREQ DATA=study1 ORDER=DATA;
      WHERE Illness=1 ;
      WEIGHT Count ;
      TABLES Test ;
      EXACT binomial ;
      TITLE1 'Sensitivity' ;  RUN ;
```

und liefert für die diagnostischen Kennzahlen des obigen Beispiels inkl. weiterer Kennzahlen wie z.B. Konfidenzgrenzen folgende Werte:

Sensitivität (*84.6%*) bzw. Anteil Falschnegative (*15.4%*),
exakte Konfidenzgrenzen (Se: *65% - 100%*),

exakter Test mit Nullhypothese H_0: Se=0.5 (*p=0.0112*), usw.

Die SAS-Statements für die Bestimmung der Kennzahl ***Spezifität*** sind entsprechend dann

```
PROC FREQ DATA=study1 ;
      WHERE Illness=0 ;
      WEIGHT Count ;
      TABLES Test ;
      EXACT binomial ;
      TITLE1 'Specificity';
      RUN ;
```

liefert folgende Werte:

Spezifität (*60 %*) bzw. Anteil Falschpositive (*40 %*),
exakte Konfidenzgrenzen (Sp: *26% - 88%*),
exakter Test mit Nullhypothese H_0: Sp=0.5 (*p=0.3770*), usw.

3 Quantitativer Test: ROC-Kurve

Warum wird eine ROC Kurve benötigt?

Zur Erinnerung: Die im vorletzten Abschnitt beschriebene Vorgehensweise vergleicht den beobachteten quantitativen Messwert eines Testes mit einem bestimmten vorgegebenen Cutoff-Wert (Schwellenwert), z.B.:

quantitativer Messwert	> Cutoff	⇒	Test +
	≤ Cutoff		Test -

Dies bedeutet, die zentrale Größe Cutoff-Level (Schwellenwert, Trennpunkt) beeinflusst stark die diagnostischen Kennzahlen wie Se, Sp, usw.. Abbildung 3.1 zeigt zwei simulierte Verteilungen. Die obere Verteilung repräsentiert beispielsweise die Nichterkrankten, die untere Verteilung die Population der Erkrankten. Die x-Achse (mit `qvalue` bezeichnet) repräsentiert die Messgröße. Je nach Wahl einer vertikalen Trennlinie werden unterschiedliche Anteile der beiden Populationen erfasst. Die Position dieser Trennlinie (=Schwellenwert, Trennpunkt oder Cutoff-Punkt) auf der x-Achse ändert demnach die Anteile für Se bzw. Sp unter den beiden Verteilungskurven.

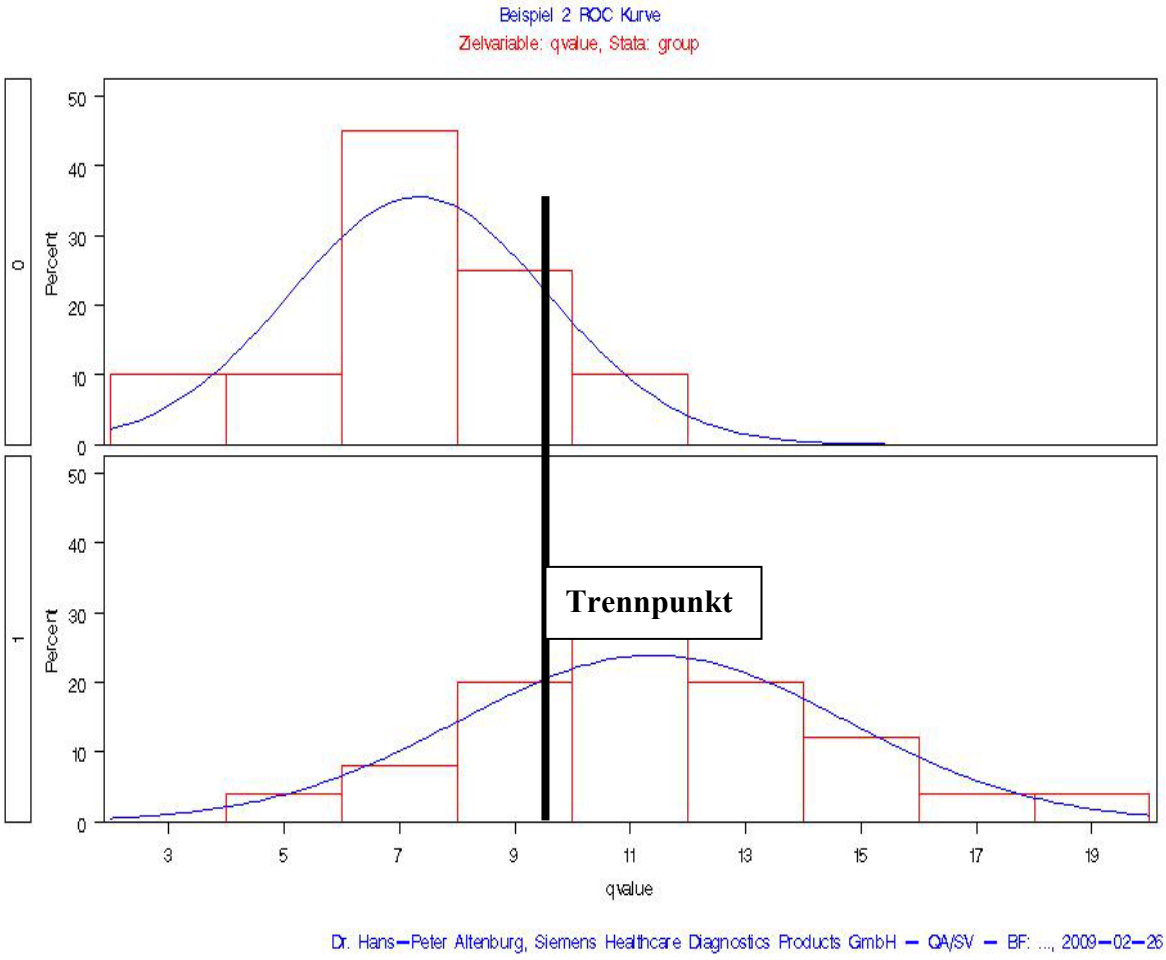

Abbildung 3.1: Lage zweier Verteilungen mit hoher Sensitivität

Jedem Cutoff-Punkt entspricht ein Punkte-Paar (1-Sp, Se), mit 1-Sp (=Anteil Falschpositiver) und Se (=Anteil Richtignegativer). Trägt man auf der Abszisse (x-Achse) den Anteil Falschpositiver (=1-Sp) und auf der Ordinate (y-Achse) die Se ab, so liefern alle Kombinationen die ROC Kurve! Abbildung 3.2 zeigt die ROC-Kurve zu obigem Beispiel.
Wie ändert sich die ROC-Kurve und die zugehörige Fläche unter der Kurve AUC (engl. **A**rea **U**nder **C**urve), wenn die Verteilungen bzw. ihre Lage zueinander sich ändern? Vier Situationen zur Verteilungs-Kurven-Separation sollen hier beschrieben werden:

- Wenig Unterschied zwischen den Verteilungen (ziemlich überlappend): Die ROC-Kurven zeigen wenig Unterschied zur Winkelhalbierenden: Die Fläche unter der Kurve AUC ist nur knapp über oder annähernd ≈ 0.5.
- Moderate Unterschiede zwischen den Verteilungen (nur noch moderat überlappend): Die ROC-Kurven liegen schon deutlich oberhalb der Winkelhalbierenden: Die Fläche unter der Kurve AUC ist nun deutlich über 0.5, z.B. AUC≈0.8.
- Gute Differenzierung zwischen den Verteilungen möglich (nur noch wenig überlappend): Die ROC-Kurven liegen schon deutlich oberhalb der Winkelhalbierenden in der linken oberen Ecke: Die Fläche unter der Kurve AUC ist nun sehr deutlich über 0.5, z.B. AUC≈0.85 bis 0.9.
- Sehr gute Differenzierung zwischen den Verteilungen möglich (nur noch kaum überlappend): Die ROC-Kurven liegen sehr deutlich oberhalb der Winkelhalbie-

renden und deutlich in der linken oberen Ecke: Die Fläche unter der Kurve AUC ist nun nahe 1, z.B. AUC≈0.98.

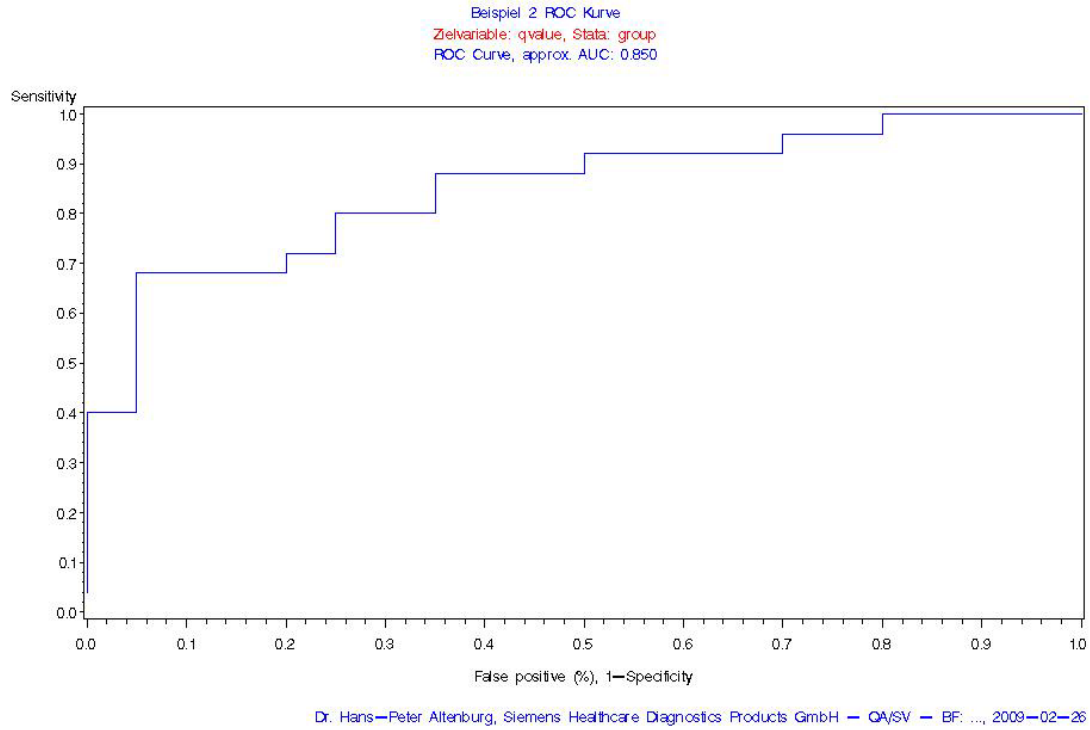

Abbildung 3.2: ROC-Kurve Datenbeispiel 2

Schlussfolgerung:

- Je näher die ROC-Kurve an der Winkelhalbierenden liegt, desto weniger kann der Test die beiden Populationen unterscheiden.
- Je mehr die Kurve in die linke obere Ecke geht, desto besser kann der Test zwischen den beiden Gruppen unterscheiden.

4 Kennzahl Fläche unter der Kurve (AUC)

Die wichtigste Kennzahl, um die Nähe zur Diagonalen zu charakterisieren ist demnach die Fläche unter der Kurve **AUC**. Je näher AUC an 0.5 liegt, desto schlechter der Test (zur Differenzierung der beiden Gruppen), und je näher AUC an 1 liegt, desto besser der Test zur Differenzierung! Der wirkliche **Vorteil,** AUC zu verwenden, liegt darin, dass die Fläche unter der Kurve einfach zu handhaben ist.
Die Fläche unter der ROC Kurve wird nicht wesentlich von der Form (Wölbung oder Schiefe) der (Häufigkeits-) Verteilungen der beiden Populationen beeinflusst. Es ist also quasi ein verteilungsfreier Ansatz!

4.1 Bestimmmung eines optimalen Trennpunktes

Die Ermittlung des optimalen Trennpunktes kann über ein logistisches Modell mit Hilfe der SAS-Prozedur LOGISTIC geschehen. Die Prozedur LOGISTIC liefert dann neben dem optimalen Trennpunkt auch noch eine Punktschätzung für die Fläche unter der Kurve AUC.

Wir gehen im Folgenden von folgender Datensituation aus: Es liegen zwei Gruppen von Messwerten vor, welche in zwei unabhängigen Populationen beobachtet wurden. Die quantitative Zielvariable wird mit der Makrovariable `&objectvar`, die Gruppenvariable mit `&byVar` bezeichnet.

Das folgende SAS-Programm passt ein logistisches Modell an, erzeugt eine SAS-Datei für die ROC-Kurve (Option `OUTROC=…`), bestimmt AUC, und erzeugt eine Datei mit den Werten des geschätzten logistischen Modells (`OUTPUT OUT=Estout …`):

```
PROC LOGISTIC DATA=&workset DESCENDING ;
MODEL &ByVar=&objectvar
   / OUTROC=ROCKurve ROCEPS=0 ;
OUTPUT OUT=EstOut P=Predicted ;
ODS OUTPUT Association=Assoc ;
RUN ;
```

AUC findet man im OUTPUT-Fenster unter den Assoziationsmassen als Parameter **c:**

Association of Predicted Probabilities and Observed Responses

Percent Concordant	99.9	Somers' D	0.998
Percent Discordant	0.1	Gamma	0.998
Percent Tied	0.0	Tau-a	0.496
Pairs	1200	**c**	**0.999 ↔ AUC**

Den optimalen Trennpunkt (Punktschätzung) erhält man dann, wenn man in der angepassten Kurve (Daten dazu in der SAS-Datei `Estout`) an der Stelle 0.5 auf der y-Achse zur x-Achse (der Skala der quantitativen Messgröße) geht (siehe Abbildung 4.1).

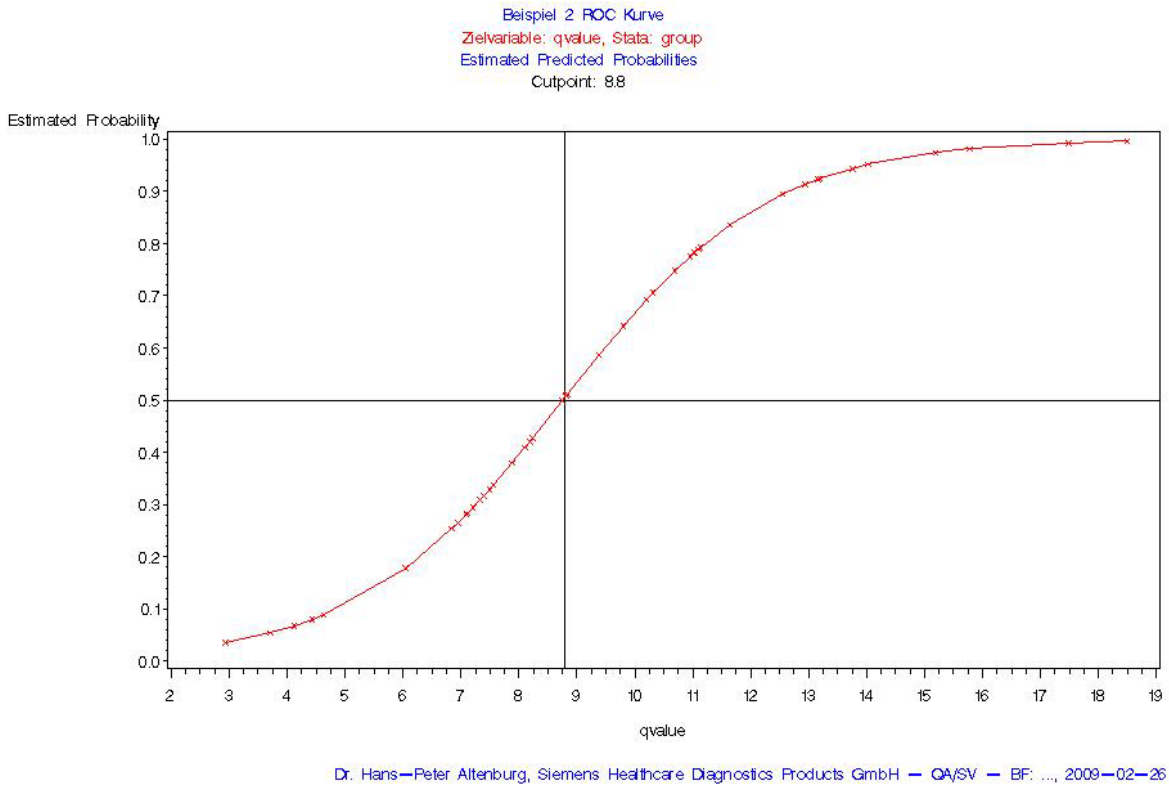

Abbildung 4.1: Angepasste logistische Funktion

Die angepasste logistische Funktion ist umso steiler (um den Cutpoint) je weniger die beiden Populationen sich überlappen, d.h. je besser sie sich trennen lassen.
Konfidenzintervalle für diesen optimalen Cutpoint (sog. Fiduzialgrenzen) lassen sich mit SAS über die SAS-Prozedur PROBIT bestimmen. Es muss dort lediglich als Modell die logistische Funktion angegeben werden, da standardmäßig in dieser Prozedur ein Probit-Modell angepasst wird.
Die ROC-Kurve erhält man dann aus den Daten der von der Prozedur LOGISTIC erzeugten Datei `ROCKurve`. Diese SAS-Datei enthält alle wichtigen Größen für die ROC-Kurve. Zur besser verständlichen Verwendung sollten die dort vorhandenen Daten / Variablen noch ein wenig aufbereitet werden, z.B. mit einem Label versehen:

```
DATA ROCKurve ;
SET ROCKurve ;
_SPECIF_=1 - _1MSPEC_ ;
_Fnegat_=1 - _SENSIT_ ;
LABEL _SPECIF_='Specificity'
      _Fnegat_='False negative (%), 1-Sensitivity'
      _1MSPEC_='False positive (%), 1-Specificity'    ;
RUN ;
```

Die Variablen _SPECIF_ bzw. _Fnegat_ werden dabei von SAS erzeugt. Die ROC-Kurve lässt sich dann einfach mit der SAS-Prozedur GPLOT darstellen:

```
PROC GPLOT DATA=ROCKurve ;
PLOT _sensit_*_1mspec_ / vaxis=0 to 1 by .1 ;
RUN ;
QUIT ;
```

Abbildung 3.2 zeigte bereits die entsprechende ROC-Kurve. In Beispiel 2 war der optimale Trennpunkt 8.8 und die AUC beträgt 0.85.

Beim SAS-Support ist ein SAS-Makro ROC-Plot erhältlich
über http://SAS "Knowledge Base" / Samples & SAS Notes im Beispiel "Sample 25017" unter der Rubrik "Downloads" herunterladen. (Zuletzt besucht am 24.06.2009.) http://Support.SAS.com\Samples_app\sample005206roc.sas.txt), bei dem über eine Option OPTIMAL eine Ausgabe erzeugt wird mit Angaben, die es ebenfalls erlauben, den optimalen Cutpoint zu bestimmen. Dort wird allerdings der Euklidische Abstand zur linken oberen Ecke als Maß verwendet. Das oben beschriebene Vorgehen über die logistische Funktion erscheint aber flexibler, da auch andere Gewichtungen von Se zu Sp gewählt werden können. Der genannte Wert 0.5 entspricht einem Verhältnis Se/Sp=0.5 (Gleichgewichtung). Es sind aber sicher auch andere Gewichtungen denkbar, wenn beispielsweise für einen neuen Test die Se eine gegenüber Sp herausragendere Rolle spielen soll und ein Wert größer als 0.5 in Frage kommt. Die ROC-Kurve ist in der Prozedur LOGISTIC in der SAS Version 9.1.3 nur experimentell verfügbar, ab SAS Version 9.2 ist sie in LOGISTIC produktiv integriert.

5 Vergleich zweier ROC Kurven bzw. zweier AUCs

Der Vergleich zweier ROC-Kurven über die AUC kann u.U. nicht einfach werden. Die Fläche kann z.B. gleich sein, aber die Kurven kreuzen sich. Liegen zwei Kurven mit ähnlicher Fläche vor, so müssen evtl. komplexere statistische Verfahren verwendet werden, wie z.B. bivariate Verfahren oder Bootstrapping, welche weiter unten besprochen werden sollen.

5.1 Vergleich zweier ROC Kurven über AUC

Häufiger Fall: Zwei diagnostische Tests mit unterschiedlichen Flächen und unterschiedlichen Fällen (es gibt dann keine Abhängigkeiten!). Das Vorgehen ist einfach. Berechne den Standardfehler der Differenz der beiden Flächen:

$$Std_{Err}(AUC_2 - AUC_1),$$

und prüfe über eine Normalverteilungsapproximation mit dem Standardfehler der Differenz. Vorsicht, dieses Verfahren kann nicht verwendet werden, wenn die gleiche Menge von Fällen für beide Teste verwendet wurde.
Der Standardfehler von AUC kann über die Prozedur FREQ ermittelt werden: AUC und Somer‘s D (ein Konkordanzmaß) hängen über die Beziehung AUC=(D+1)/2 voneinander ab. Das folgende SAS-Programm zeigt, wie man den Wert von Somer’s D bestimmt. Die Variablenbezeichnungen sind wie oben bereits angegeben.

```
PROC FREQ DATA= &workset ;
TABLES &objectvar * &ByVar / NOPRINT MEASURES ;
RUN ;
```

Im SAS-Output findet man D unter der Bezeichnung `Somer‘s D R|C` (die Werte hier für das oben bereits verwendete Datenbeispiel 2):

```
Statistic                                Value       ASE
 . . .
Somers' D C|R                           0.3535    0.0576
Somers' D R|C                           0.7000    0.1116 ←
```

Hieraus ergeben sich die Schätzwerte für AUC und den Standardfehler: Std_{Err}(AUC)= Std_{Err} (D) und das approximative 95% Konfidenzintervall:

$\Rightarrow$ AUC = (1+0.7)/2 = 0.85,
95%-Konfidenzintervall: 0.74 / 0.96

5.2 Binormales Modell für Modellierung und den Vergleich von ROC’s

Eine aus statistischer Sicht deutlich aufwändigere Lösung ist die Modellierung oder Anpassung der ROC-Kurve an ein Binormal-Modell.

Binormal-Modell:

Wir gehen von einer normal-verteilten Meßgröße aus:

Verteilung **Erkranke** $\sim N(\mu_1, \sigma_1)$
Verteilung **Nichterkrankte** $\sim N(\mu_0, \sigma_0)$

Die funktionale Form der ROC-Kurve lässt sich durch

$$G(t)=\Phi(a+b\ \Phi^{1}(x)),$$

beschreiben, wobei

$$a=(\mu_1 - \mu_0) / \sigma_1$$

und

$$b= \sigma_0/\sigma_1 .$$

Für AUC gilt dann: $AUC=\Phi(a/\sqrt{(1+b^2)})$

Die Schätzung bzw. Anpassung der ROC-Kurve an ein Binormal-Modell kann dann über die SAS-Prozedur NLMIXED erfolgen:

SAS Programm-Code:

```
PROC NLMIXED DATA=&workset ;
  PARAMETERS m1=0 m0=0 s1=1 s0=1;
  IF &ByVar=1  THEN m=m1;
          ELSE IF &ByVar=0 THEN m=m0;
  IF &ByVar=1  THEN s=s1*s1;
          ELSE IF &ByVar=0 THEN s=s0*s0;
  a=(m1-m0)/s1;
  b=s0/s1;
  MODEL &objectVar ~ normal(m,s);
  ESTIMATE 'a' a;
  ESTIMATE 'b' b;
  ESTIMATE 'AUC' probnorm(a/sqrt(1+b**2));
RUN ;
```

Für das Datenbeispiel 2 liefert der OUTPUT folgende Schätzwerte:

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
a	**1.2368**	0.3048	45	4.06	0.0002	0.05	0.6230	1.8506
b	**0.6676**	0.1416	45	4.71	<.0001	0.05	0.3824	0.9528
AUC	**0.8482**	0.05548	45	15.29	<.0001	0.05	0.7364	0.9599

Von besonderem Interesse in diesem OUPUT sind die ersten drei Spalten, welche die Parameterschätzungen bzw. zugehörigen Standardfehler repräsentieren und die beiden letzten Spalten, welche die untere und obere Grenze des entsprechenden Konfidenzintervalls des jeweiligen Parameters wiedergeben.

SAS Programm-Code für den Plot der geschätzten ROC-Kurve:

```
DATA ROCData;
  SET ROCKurve;
  _bsensit_=probnorm(1.24+0.67*probit(_1mspec_));
RUN;

axis1 length=12cm order=0 to 1 by 0.2 label=(f=swissb h=2)
```

```
      value=(font=swissb h=2);
axis2 length=12cm order=0 to 1 by 0.2 label=(a=90 f=swissb h=2)
      value=(font=swissb h=2);
symbol1 v=none i=join w=3 l=1 c=black;
symbol2 v=none i=join w=3 l=3 c=black;

          PROC GPLOT DATA=ROCData;
            PLOT (_sensit_ _bsensit_)*_1mspec_
           / haxis=axis1 vaxis=axis2 overlay;
          RUN;
          QUIT;
```

Die folgende Abbildung 5.2.1 zeigt die beobachtete und die geschätzte ROC-Kurve.

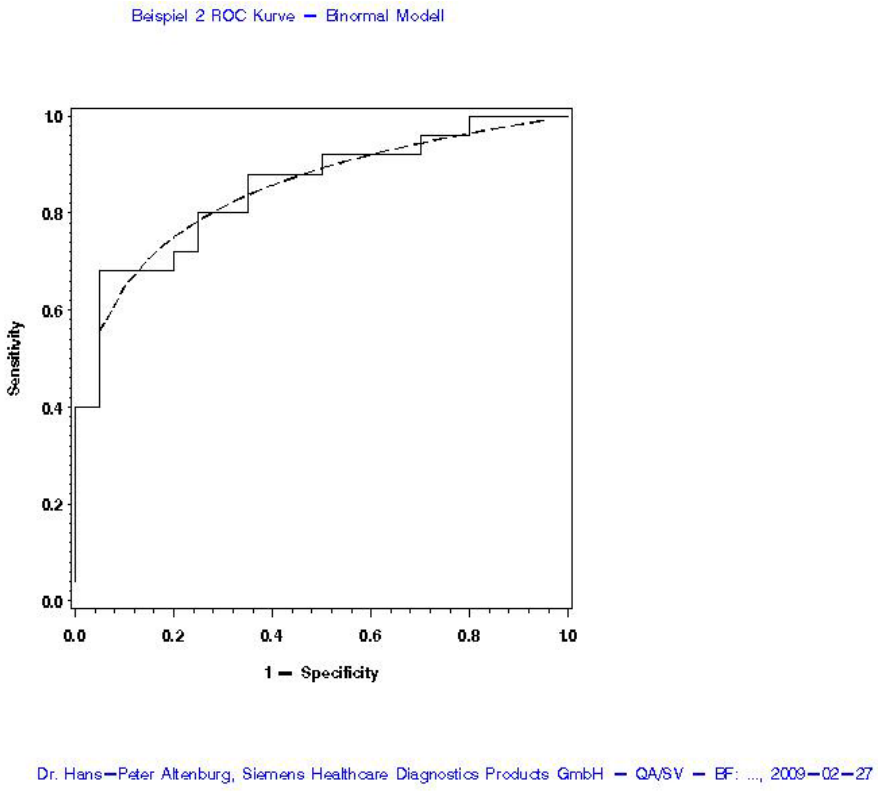

Abbildung 5.2.1: ROC-Kurve Datenbeispiel 2 beobachtet vs. geschätzt

5.3 Direkte Binormal-Schätzung

Neben der Anpassung an ein binormales Modell kann eine ROC-Kurve auch direkt über die vorliegenden Daten geschätzt werden. Zwei Schritte sind dafür erforderlich:

- Probit-Transformation der beiden Variablen der ROC-Kurve (zur Linearisierung),
- Prozedur REG zur Schätzung der Konstanten a und b.

SAS Programm-Code direkte Binormal-Schätzung:

```
DATA RocKurve2 ;  /* Probit-Transformation */
SET ROCKURVE ;
_y=PROBIT(_sensit_) ;
_x=PROBIT(_1mspec_) ;
RUN ;

PROC REG DATA=RocKurve2 ;
MODEL _y=_x ;
RUN ;
```

Für das Datenbeispiel 2 erhält man folgende Schätzwerte:

Direkte Methode:	**a=1.23**	**b=0.69**
NLMixed:	**a=1.24**	**b=0.67**

Die folgende Abbildung 5.3.1 zeigt beide geschätzte Kurven.

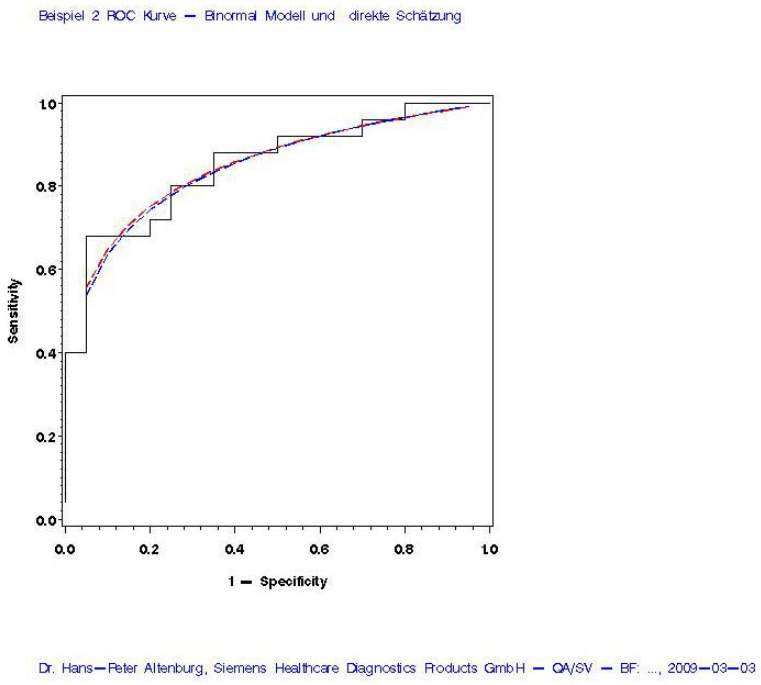

Abbildung 5.3.1: ROC-Kurven geschätzt direkt und Binormal-Modell

In diesem Beispiel unterscheiden sich beide Kurven kaum. Auf den ersten Blick erscheint dies als attraktive Möglichkeit. Jedoch ist dieses Vorgehen nicht geeignet für Inferenzbetrachtungen. Der Standardfehler (aus der Prozedur REG) unterschätzt den wahren Wert. Außerdem sind die Werte nicht unabhängig (sie stammen aus einer kumulierten Verteilung). Deshalb sollte die direkte Schätzung nur für eine Punktschätzung aber nicht für Inferenzaussagen verwendet werden!

Das Binormal-Modell ist flexibler und erlaubt auch den Vergleich mehrerer AUC's, den Vergleich mehrerer Modelle oder die Adjustierung nach verschiedenen Kovariablen.

5.4 Konfidenzintervall für AUC über Bootstrap-Schätzung

Bootstrapping ist eine nichtparametrische Methode zur Schätzung von Variabilität oder Verzerrungen von Parameterschätzungen. Sie ist z.B. besonders in folgenden Situationen eine gut geeignete Vorgehensweise:

- es ist wenig über die Parameterschätzung bekannt,
- die Parameterschätzung ist mathematisch nur schwer zu behandeln, oder
- Verteilungsannahmen für die Parameterschätzung sind nur schwer zugänglich.

Speziell im Fall von ROC-Kurven-Analysen kann sie verwendet werden, um nichtparametrische Konfidenzintervalle, Hypothesentests zum Vergleich von ROC-Kurven, zum Ausgleich zu „optimistischer" multivariater Vorhersagemodelle über Logistische Regression zu bestimmen.
Das Prinzip des Bootstrappings ist einfach. Über Zufalls-Subsamples und deren Einzelanalysen werden die Daten zusammengefasst und Perzentile der in Frage kommenden Größe bestimmt. Zentrales Element für die Verwendung in SAS ist dabei die Prozedur SURVEYSELECT zur Auswahl der Zufallsstichproben. Der folgende Beispiel-Programm-Code zeigt ein Grundmuster, das in vielen Fällen verwendet werden kann. Verschiedene Sampling-Methoden (`METHOD= …`) können je nach Aufgabe verwendet werden (z.B. Ziehen mit oder ohne Zurücklegen usw.). Jede Einzelstichprobe wird dann in einer SAS-Datei (`OUT=…`) abgelegt und kann dann entsprechend weiter analysiert werden. Die anschliessende Analyse erfolgt z.B. mit Hilfe der Prozedur UNIVARIATE. In der Regel genügen 3000 bis 5000 (`REP= …`) Wiederholungen, um stabile Ergebnisse zu erhalten.

SAS-Programm-Code Bootstrapping (Beispiel für Subsamples):

```
PROC SURVEYSELECT DATA=&dset  /* Data Set              */
     METHOD=urs               /* Unrestricted Sampling*/
     N=&n                     /* Umfang Subsample     */
     OUT=bootsample           /* Ausgabe data set     */
     REP=&Wdh                 /* Anzahl Wdh.          */
     NOPRINT;
RUN ;
```

Als Beispiel wählen wir die Bestimmung eines Konfidenzintervalls für die AUC mit den Daten des hier bisher bereits mehrfach verwendeten Datenbeispiels 2.

Beispiel: 95%-Konfidenzintervall **AUC** (Datenbeispiel 2)
Zum Vergleich werden nochmals die entsprechenden Werte der beiden anderen Ansätze über Somer's D und NLMIXED angegeben.

AUC-Punktschätzung:	0.85	
95%-Konfidenzintervall:	Somers‘ D:	0.74 / 0.96
	NLMIXED:	0.74 / 0.96
	Bootstrap:	0.72 / 0.95

6 Fehlerquellen

Zahlreiche Fehlerquellen können sehr negative Effekte auf die diagnostische Wertigkeit eines neuen Testes aufzeigen. Hier sollen nur einige wenige aufgelistet werden:

- Effekt von „random noise” (zufälliges Rauschen / weißes Rauschen) kann die Kurve durch zufällige Variationen beeinflussen,
- z.B. korreliert ein neuer Test perfekt mit dem Gold-Standard ($\Rightarrow$AUC=1), hinzufügen von weißem Rauschen (Misklassifikationen) verringert dann u.U. den Wert von AUC,
- weißes Rauschen kann zu falschen Schlussfolgerungen führen, wie etwa „*neuer Test ist besser*“, in Wirklichkeit ist er es aber doch nicht.
- Unabhängigkeit vom Gold-Standard: Probleme können entstehen wenn der Test und der verwendete Gold-Standard nicht unabhängig sind, dies führt dann z.B. zu einem großen Wert von AUC.
- Der Gold-Standard ist von schlechter Qualität: Das Rauschen hat dann einen Effekt, was zu einer Unterschätzung der Parameter (sog. „nicht differentielle Fehlklassifikation”) führen kann.

Literatur

[1] DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837-845

[2] Gönen, M. (2007): Analyzing Receiver Operating Characteristic Curves with SAS. Cary NC: SAS Publishing Series

[3] Pepe, M.S. (2003): The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press.

[4] Tosteson, A.N. and Begg, C.B. (1988): A general regression methodology for ROC curve estimation. Medical Decision Making 8, 204-215