Simulationen und das Problem der nachträglichen Kategorisierung am Beispiel der AUC

Stefan Beimel
PAREXEL International GmbH
Am Bahnhof Westend 15
14059 Berlin
stefan.beimel@PAREXEL.com

Zusammenfassung

Bei diagnostischen Studien spielt die Area Under the Receiver Operating Characteristic Curve (AUC) eine herausragende Rolle. Die Simulation einer AUC erfordert, dass zwei Verteilungen in einer ganz bestimmten Beziehung zueinander stehen. Diese Beziehung ist einfach darzustellen, wenn es sich um zwei Normalverteilungen handelt.
Sind die Zufallsvariablen diskret, liegt es nahe, zunächst entsprechende normalverteilte Zufallszahlen zu erzeugen und diese dann geeignet den diskreten Werten zuzuordnen. Diese Zuordnung ist einfach zu programmieren - führt aber leider zu falschen Ergebnissen. Der Beitrag stellt eine Möglichkeit dar, wie trotz nachträglicher "Diskretisierung" die AUC entsprechend den Anforderungen simuliert werden kann.

Schlüsselwörter: AUC, Simulation, Kategorisierung

1 Die Aufgabe

In diagnostischen Studien wird untersucht, ob die Ergebnisse einer diagnostischen Methode geeignet sind, zwischen gesunden und kranken Personen zu unterscheiden. Das Ergebnis einer Methode kann dichotom (z. B. gesund vs. krank), ordinal (z. B. gemessen anhand einer Skala) oder metrisch sein (z. B. Laborwerte) (Tabelle 1).

Tabelle 1: Messergebnisse in einer diagnostischen Studie

Einteilung nach Goldstandard (“wahrer Status”)	Patient	Ergebnis der diagnostischen Methode
gesund	1	x_{01}
	…	…
	n_0	x_{0n_0}
krank	1	x_{11}
	…	…
	n_1	x_{1n_1}

Es gibt verschiedene Maßzahlen, um die Güte einer diagnostischen Methode einzuschätzen:

Sensitivität: Die Wahrscheinlichkeit, dass ein Kranker richtigerweise als krank erkannt wird (true positive)

Spezifität: Die Wahrscheinlichkeit, dass ein Gesunder richtigerweise als gesund erkannt wird (true negative)

AUC: Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Gesunder einen niedrigeren Wert als ein zufällig ausgewählter Kranker hat, wobei o.B.d.A. angenommen werden kann, dass höhere Werte mit kränkeren Patienten assoziiert sind.

$$AUC = P(X_0 < X_1) + \frac{1}{2} P(X_0 = X_1)$$

Die Bezeichnung AUC (Area Under the Receiver Operating Characteristic Curve) rührt daher, dass diese Maßzahl auch graphisch mit Hilfe von Sensitivität und Spezifität interpretiert werden kann. Der Korrekturterm $\frac{1}{2} P(X_0 = X_1)$ muss bei ordinaler Skalierung der Ergebnisse berücksichtigt werden.

Es sollte untersucht werden, ob das Skalenniveau der Ergebnisse bei gleichbleibender AUC einen Einfluss auf die statistischen Eigenschaften hat. Dazu sollten Simulationen durchgeführt werden.

2 Die Simulation

Es gibt viele Möglichkeiten, zwei Verteilungen für X_0 und X_1 zu finden, um eine bestimmte AUC zu simulieren.
Im Fall, wenn X_0 normalverteilt ist mit $N(\mu_0, \sigma_0^2)$, und $X_1 \sim N(\mu_1, \sigma_1^2)$ ergibt sich

$$AUC = P(X_0 < X_1) = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right)$$

Bei gegebener AUC, gegebenem σ_0^2 und σ_1^2, und unter der (vereinfachenden, aber nicht einschränkenden) Annahme, dass $\mu_0 = 0$ ist, kann μ_1 berechnet werden als

$$\mu_1 = \Phi^{-1}(AUC)\sqrt{\sigma_0^2 + \sigma_1^2}$$

Nun werden die Werte dieser kontinuierlichen Verteilungen noch in Skalenwerte umgewandelt. Die Bereiche der Kategorisierung (siehe Abbildung 1) haben im Beispiel konstante Breite und sind symmetrisch um $\mu_1/2$ angeordnet.

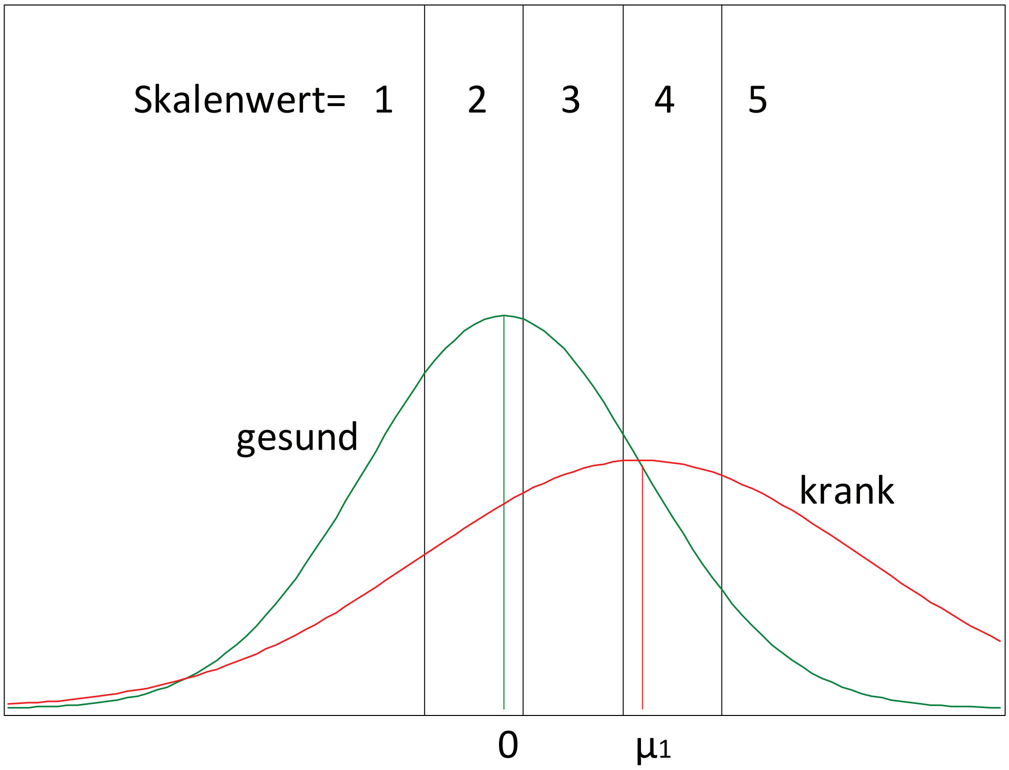

Abbildung 1: Zwei Normalverteilungen für die Simulation von Werten einer AUC, und die Grenzen für die Einteilung dieser simulierter Werte in Kategorien am Beispiel einer 5-Punkte-Skala

3 Das Problem

Während die Ergebnisse der Simulation der Normalverteilungen im Mittel tatsächlich die theoretisch erwartete AUC ergab, wich die AUC nach der Kategorisierung abhängig vom Skalenniveau teilweise erheblich von den erwarteten Werten ab. Diese Abweichung lässt sich berechnen. Aus den Grenzen der Skalenwerte, den Parametern der Verteilungen ergibt sich am Beispiel einer Skala mit 5 Ausprägungen (siehe Abbildung 1):

$$
\begin{aligned}
AUC &= P(X_0 < X_1) + \frac{1}{2}P(X_0 = X_1) \\
&= P(X_0 = 1, X_1 = 2) + P(X_0 = 1, X_1 = 3) + \cdots + P(X_0 = 4, X_1 = 5) + \\
&\quad \frac{1}{2}\big(P(X_0 = 1, X_1 = 1) + \cdots + P(X_0 = 5, X_1 = 5)\big)
\end{aligned}
$$

Da X_0 und X_1 unabhängig voneinander sind, kann man schreiben:

$$P(X_0 = i, X_1 = j) = P(X_0 = i)\, P(X_1 = j) \qquad i, j = 1, \dots, 5$$

Nun lässt sich für jede Ausprägung der Skalen die Wahrscheinlichkeit berechnen:

$$P(X_s = i) = \Phi\left(\frac{\begin{matrix}obere\\Grenze\end{matrix} - \mu_s}{\sigma_s}\right) - \Phi\left(\frac{\begin{matrix}untere\\Grenze\end{matrix} - \mu_s}{\sigma_s}\right) \qquad s = 0,1; \quad i, j = 1, \dots, 5$$

Abbildung 2 stellt an einem Beispiel die Abweichung der kategorisierten AUC von der ursprünglich erwarteten AUC für mehrere Skalenniveaus dar. Bei einer erwarteten AUC von 0,8 beträgt der Unterschied bei einer Variablen mit 2 Ausprägungen beispielsweise 10%! Damit ist die ursprüngliche Fragestellung („Hat das Skalenniveau bei gleichbleibender AUC einen Einfluss auf die statistischen Eigenschaften?“) nicht mehr zu beantworten.

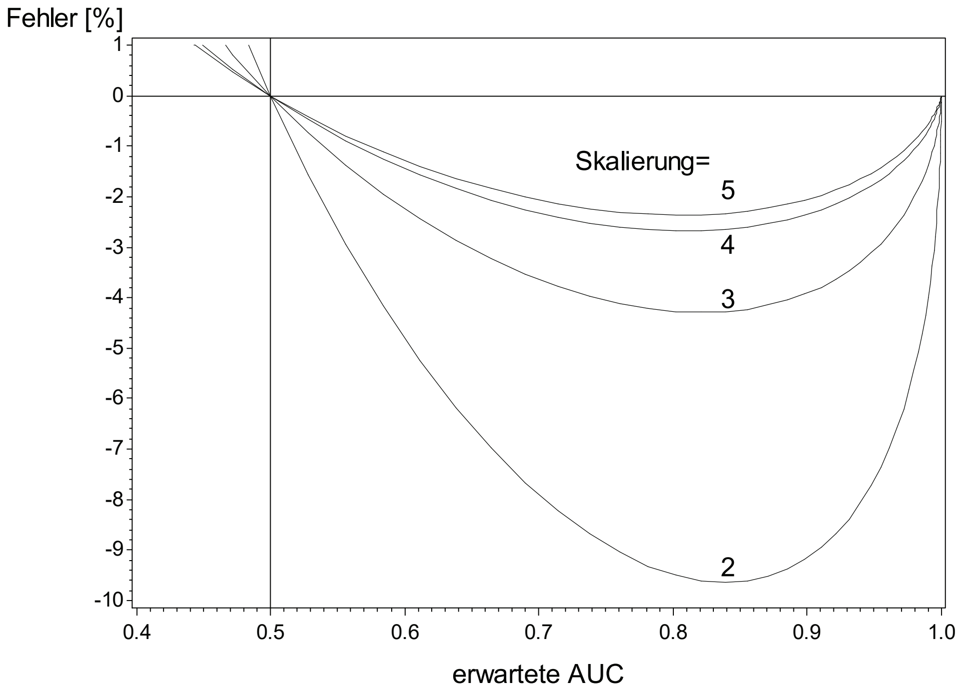

Abbildung 2: Abweichung der tatsächlichen AUC von der erwarteten AUC am Beispiel von verschiedenen Skalenniveaus

4 Die Lösung

Es ist trotzdem möglich, die AUC über die Normalverteilungen zu simulieren. Allerdings muss nun das μ_1 anders ermittelt werden als mit der Formel aus Kapitel 2. Stattdessen wurde das Bisektionsverfahrens angewandt.

Startpunkt sind zwei Werte (μ_{low} und μ_{high}, siehe Abbildung 3), die ganz sicher kleiner bzw. größer als das gesuchte μ_1 sind. Dann wird angenommen, μ_1 wäre genau der Mittelwert von μ_{low} und μ_{high}, und die AUC wird entsprechend den Formeln in Kapitel 3 berechnet. Ist die AUC größer als erwartet, wird μ_{high} auf den Wert μ_1 gesetzt; ist die AUC zu klein (wie im ersten Schritt in Abbildung 3), wird μ_{low} auf den Wert von μ_1 gesetzt. Damit hat sich das Suchintervall halbiert. Ist der Betrag der Differenz von berechneter AUC und erwarteter AUC unter einer vorher festgelegten Genauigkeitsgrenze ε, kann der Algorithmus gestoppt werden.

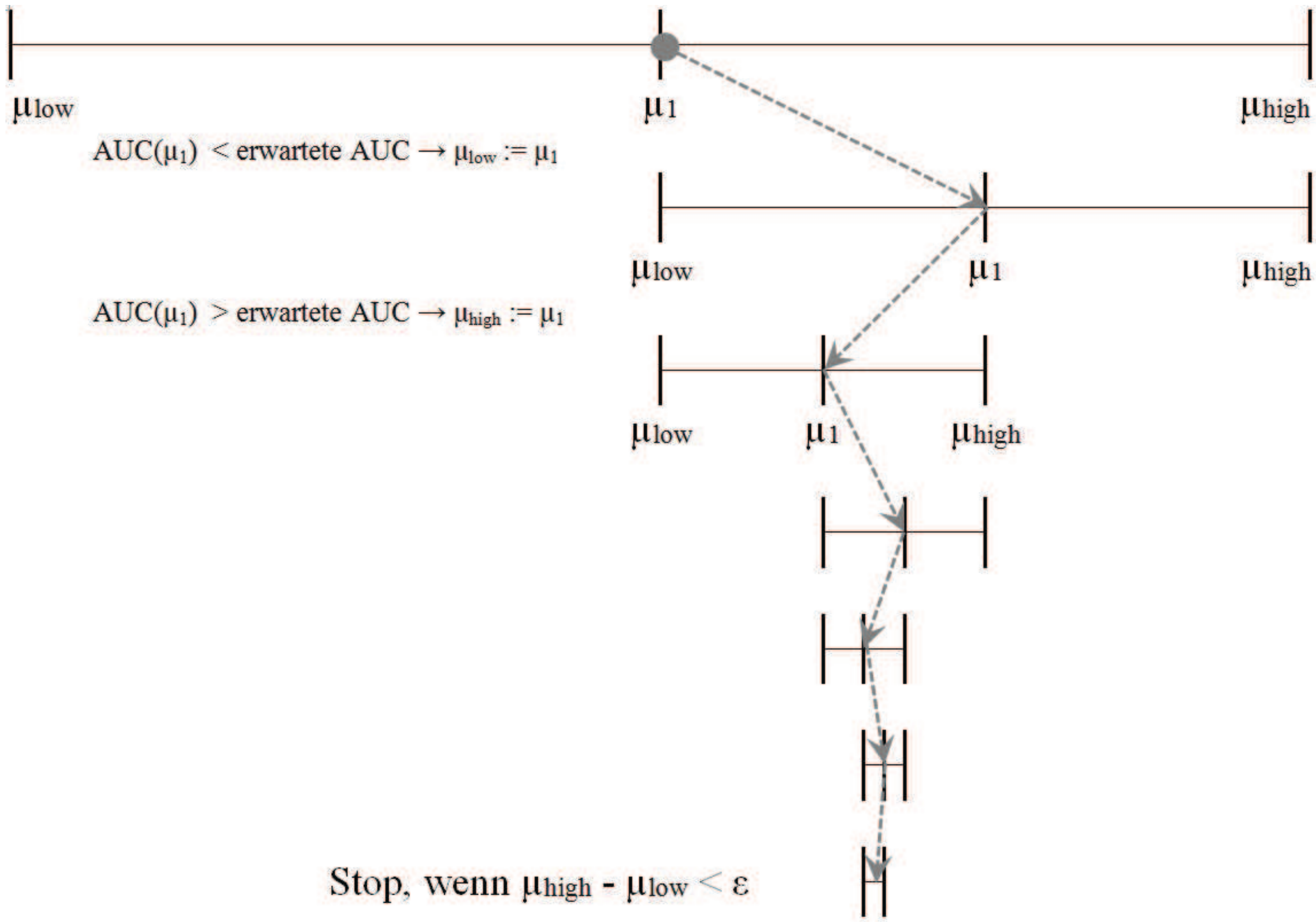

Abbildung 3: Bisektionsalgorithmus zur Bestimmung von μ_1

Dieses Verfahren ist sehr robust und liefert für alle Einstellungen die richtige AUC.