Vergleich der Wilson- und Wald-KI

Der Grund, wieso die Überdeckungswahrscheinlichkeit beim Wald-KI in der Nähe von 0 und 1 deutlich schlechter ist als die vorgegebene Sicherheitswahrscheinlichkeit $1-\alpha$ ist nicht in erster Linie darin zu suchen, dass es zu h symmetrisch ist. Der Grund hierfür lässt sich aufzeigen, wenn von den beiden Lösungen p_- und p_+ der zugrundeliegenden quadratischen Gleichung $(h-p)^2 = \frac{z^2}{n} \cdot p \cdot (1-p)$ aus argumentiert wird.

Zuerst ein paar Abkürzungen: $z := z_{1-\alpha/2}$; $a := \frac{1}{1+\frac{z^2}{n}}$; p_m : Mittelwert des Wilson-KI

Mit diesen Abkürzungen folgt:

1. Der Mittelwert eines Wilson-KI ist das gewichtete Mittel aus h und $\frac{1}{2}$.

   $$p_m = a \cdot h + (1-a) \cdot \frac{1}{2}$$

2. Die Lage des Wald-KI ist im Vergleich zum Wilson-KI zu den Rändern verschoben.

   $$\frac{1}{2} - p_m = a \cdot \left(\frac{1}{2} - h\right)$$

3. Das Wald-KI liegt nur dann im Intervall $\left[0;1\right]$, falls gilt:

   $$h \in \left[1-a; a\right]$$

4. Die Länge des Wilson-KI ist genau dann größer als die Länge des Wald-KI, falls gilt:

   $$\left|h - \frac{1}{2}\right| > \frac{1}{2} \cdot \sqrt{1 - \frac{1}{2 + z^2/n}}$$

Aus diesen Gründen ist die Überdeckungswahrscheinlichkeit für p –Werte nahe 0 oder 1 deutlich kleiner als die Vorgabe.

Das kann man mit Simulationen noch untermauern.

Übrigens, wenn man „ehrlich“ ist und von der folgenden Lösung ausgeht

$$p_{+,-} = a \cdot \left(h + \frac{z^2}{2n} \pm z \cdot \sqrt{\frac{h \cdot (1-h)}{n} + \left(\frac{z}{2n}\right)^2} \right), \quad (1)$$

muss man nicht die unsinnige Aussage machen „ich ersetze das p auf der rechten Seite der Gleichung von $(h-p)^2 = \frac{z^2}{n} \cdot p \cdot (1-p)$

durch h, auf der linken Seite aber nicht.

Betrachtet man wirklich die Lösung (1), so erkennt man, dass hier $\frac{z^2}{n}$ vernachlässigt werden kann – und damit auch $\frac{z^2}{2n}$ und $\frac{z^2}{4n^2}$. Der Term a wird durch 1 abgeschätzt.

Dann erst entsteht $p_{+,-} \approx h \pm z \cdot \sqrt{\frac{h \cdot (1-h)}{n}}$.
So – und nun kann man diesen Term stochastisch interpretieren – vorher nicht.
Vorher wurde Algebra betrieben.
Nun kann $\sqrt{\frac{h \cdot (1-h)}{n}}$ interpretiert werden als Standardfehler oder als Abschätzung der unbekannten Varianz durch die Varianz von h.
Algebra trifft Stochastik – eigentlich gut!