Zur Gefahr, stetige Zufallsvariablem zu dichotomisieren[1]

OLIVER KUß, DÜSSELDORF

[1] Das Original erschien in Teaching Statistics (Volume 35, Number 2, Summer 2013; S. 78–79). Originaltitel: The danger of dichotomizing continuous variables: A visualization
Übersetzung, Bearbeitung und Kürzung: J. MEYER

Zusammenfassung: *Fünf sehr verschiedene Streudiagramme (Scatterplots) werden vorgestellt, die jeweils zur gleichen Vierfeldertafel führen, wenn man die Zufallsvariablen dichotomisiert. Aus den identischen Vierfeldertafeln können falsche Schlüsse gezogen werden.*

In der Statistik geht es darum, die Komplexität von Daten zu reduzieren und den dabei entstehenden Informationsverlust so klein wie möglich zu halten. Mitunter kann die Komplexitätsreduktion aber auch dazu führen, wichtige in den Daten enthaltene Informationen zu verschleiern oder auszublenden.

In der Medizin dichotomisieren (d. h. auf zwei Kategorien reduzieren) Forscher gerne stetige Zufallsvariablen mit dem Ziel, die Analysen und die Interpretation der Befunde zu vereinfachen. Häufig werden jeweils zwei Kategorien als angemessen empfunden, etwa „normal“ / „anomal“ oder „behandeln“ / „nicht behandeln“. Warnungen gegen diese Praxis werden im allgemeinen kaum oder gar nicht beachtet.

1973 gab Anscombe vier Streudiagramme an, die sehr unterschiedliche Beziehungen zwischen den Zufallsgrößen X und Y ausdrückten. Gleichwohl hatten X und Y jeweils gleiche Erwartungswerte und Varianzen. Zudem führten alle vier Streudiagramme zur gleichen Ausgleichsgeraden und Korrelationskoeffizienten.

Dies hat mich angeregt, fünf Streudiagramme anzugeben, die nach Dichotomisierung jeweils die gleiche Vierfeldertafel ergeben:

	Y positiv	Y negativ
X positiv	25	25
X negativ	25	25

Gemäß der Vierfeldertafel gibt es keine Beziehung zwischen X und Y.

Das ist für das erste Streudiagramm (Abb. 1) auch richtig, nicht aber für Abb. 2, das einen quadratischen Zusammenhang zeigt, und auch nicht für Abb. 3, das eine sinusförmige Beziehung aufweist. In Abb. 4 findet man eine Heteroskedastizität (also eine unterschiedliche Streuung der Daten in verschiedenen Bereichen; vgl. Titelbild), und Abb. 5 zeigt eine Mischung zweier Populationen. Bei allen Streudiagrammen ist die Rechtsachse die X-Achse, und die Hochachse ist die Y-Achse.

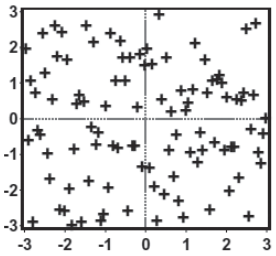

Abb. 1: X und Y haben keinen erkennbaren Zusammenhang

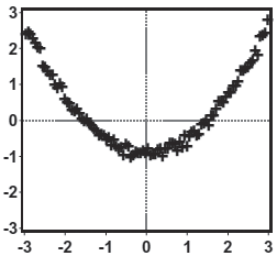

Abb. 2: Y hängt quadratisch von X ab

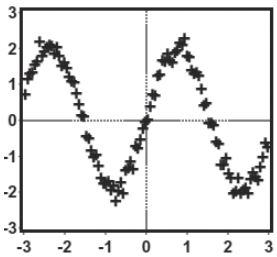

Abb. 3: Y hängt sinusförmig von X ab

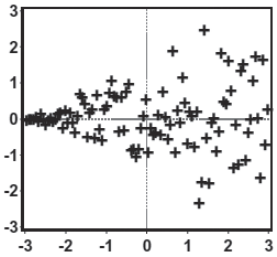

Abb. 4: Die Streuung nimmt nach rechts zu

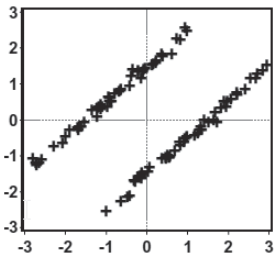

Abb. 5: Zwei Populationen

Literatur

Anscombe, F. J. (1973): Graphs in Statistical Analysis. In: *American Statistican* 27 (1), S. 17–21.

Anschrift des Verfassers

Oliver Kuß
Institut für Biometrie und Epidemiologie
Deutsches Diabetes-Zentrum, Leibnitz-Zentrum für Diabetes-Forschung an der
Heinrich-Heine-Universität Düsseldorf
oliver.kuss@ddz.uni-duesseldorf.de