Markus VOGEL, Heidelberg

„Stochastik reloaded“ – Mit Daten und Zufall in die Unterrichtspraxis

„Daten“ und „Zufall“ bilden zusammen eine Leitidee der Bildungsstandards. Mit der curricularen Zusammenführung der empirischen Welt der Daten und der wahrscheinlichkeitstheoretischen Welt von zufälligen Vorgängen stellt sich die Frage, wie man im Unterricht vorgehen kann, so dass die Idee, die beiden Welten zusammenzuführen, auch unterrichtspraktisch konkret wird. Im Folgenden werden dazu wesentliche stochastikdidaktische Überlegungen und exemplarisch unterrichtpraktische Konkretisierungen vorgestellt.

1. Eine Welt voller Daten

Wir leben in einer Welt voller Daten. Daten sind omnipräsent – schon beim einfachen Aufschlagen einer Zeitung oder beim Einschalten der Nachrichten begegnen uns Daten und verlangen von uns Situationen und damit verbundene Sachverhalte zu prüfen, abzuwägen und Entscheidungen zu treffen. Durch Daten – als Zahlen mit Kontext – werden Phänomene, Situationen oder, ganz allgemein gesprochen, Informationen der natürlichen, technischen oder sozialen Umwelt quantifiziert. Gerade wenn es um wichtige Entscheidungen geht, sollten diese Entscheidungen nicht rein subjektiver Willkür unterworfen, sondern möglichst objektiv und allgemein nachvollziehbar auf der Basis von guten Daten begründet sein. Hierin begründen sich die Anstrengungen einer sorgfältigen Datenanalyse. Die Aufgabe besteht im Kern immer darin, im Überangebot aller Dateninformationen, bildlich gesprochen im Nebel der Datenwolke, Strukturen ausfindig zu machen und diese als Datentrend zu beschreiben.

Wenn es darum geht, dass die Schülerinnen und Schüler lernen, „technische, natürliche, soziale und kulturelle Erscheinungen und Vorgänge mit Hilfe der Mathematik wahr[zu]nehmen, [zu] verstehen und unter Nutzung mathematischer Gesichtspunkte [zu] beurteilen“ (KMK, 2003, S. 6), wird offensichtlich, dass die Leitidee Daten und Zufall hinsichtlich der zugrunde liegenden anwendungsorientierten Perspektive einen wesentlichen Beitrag dazu leisten kann, um diesen schulischen Anspruch einlösen zu helfen. Darüber hinaus geht es in einer strukturmathematischen Perspektive auch um die Handhabung der verwendeten mathematischen Werkzeuge (z. B. Funktionen). Durch den kontext- und zweckgebundenen Gebrauch können sich diese in ihrer mathematischen Begrifflichkeit weiter erschließen (vgl. Vollrath, 1989).

2. Daten und Zufall

In den KMK-Bildungsstandards für den mittleren Schulabschluss aus dem Jahr 2003 (KMK, 2003, S. 12) heißt es:

„Die Schülerinnen und Schüler

- werten graphische Darstellungen und Tabellen von statistischen Erhebungen aus,
- planen statistische Erhebungen,
- sammeln systematisch Daten, erfassen sie in Tabellen und stellen sie graphisch dar, auch unter Verwendung geeigneter Hilfsmittel (wie Software),
- interpretieren Daten unter Verwendung von Kenngrößen,
- reflektieren und bewerten Argumente, die auf einer Datenanalyse basieren,
- beschreiben Zufallserscheinungen in alltäglichen Situationen,
- bestimmen Wahrscheinlichkeiten bei Zufallsexperimenten."

Zusammenfassend lässt sich sagen, dass es grundsätzlich darum geht, dass die Schülerinnen und Schüler lernen, Fragen an empirische Phänomene ihrer erlebten Umwelt zu stellen und mit den zur Verfügung stehenden elementaren mathematischen Mitteln zu beantworten. Die Daten sind der Wahrscheinlichkeitsrechnung vorgeordnet und der statistische Aspekt geht über das bloße Erstellen von Grafiken als Teil des Sachrechnens deutlich hinaus. Der Zufall wird ebenfalls empirisch erfahrbar über die phänomenologische Wahrnehmung und experimentelle Untersuchungen eingebunden.

Wenn die Daten unmittelbar naheliegende Anknüpfungspunkte für den Mathematikunterricht im Sinne der Leitidee Daten und Zufall bieten, wie und wo sind dann Überlegungen zur Wahrscheinlichkeitsrechnung anzustellen, was haben eigentlich Daten und Zufall miteinander zu tun? Wenn diese Frage (insbesondere angesichts der derzeitigen Diskussion um die KMK-Bildungsstandards für die Sekundarstufe II) keine explizite Antwort, auch in der curricularen Verankerung, findet, dann besteht die Gefahr, bereits früher als unzulänglich erkannten Entwicklungen wieder den Weg zu bereiten: „[…] daß [sic!] in den meisten […] Mathematikwerke[n] statistische und wahrscheinlichkeitstheoretische Passagen nahezu beziehungslos (und zuweilen mit falschem Bezug) hintereinandergesetzt [sic!] sind." (Schupp, 1982, S. 214 mit Verweis auf Steinbring, H. & Strässer, R. (Hrsg.), 1981)

Daten und ihre Analyse lassen sich hinsichtlich ihrer Ausrichtung in zwei Bereiche unterteilen:

1. Rückblick und Bestandsaufnahme: Die Datenanalyse umfasst die Beschreibung eines in statistischen Daten erfassten Ist-Zustands der Realität. Mit dem Blick auf einen Jetzt-Zustand (und dem damit verbundenen Blick in die Vergangenheit, aus der dieser Jetzt-Zustand hervorgeht,) ist

sie in der Gegenwart verankert. Die Daten bilden nicht die Realität in ihrer Komplexität ab, sondern nur ein vereinfacht abbildendes Modell der Realität (z. B. Daten einer Volkszählung). Die inhaltliche Perspektive dieses Modells ergibt sich aus der vorgeordneten erkenntnisleitenden Fragestellung, die zur Datenerhebung geführt hat.

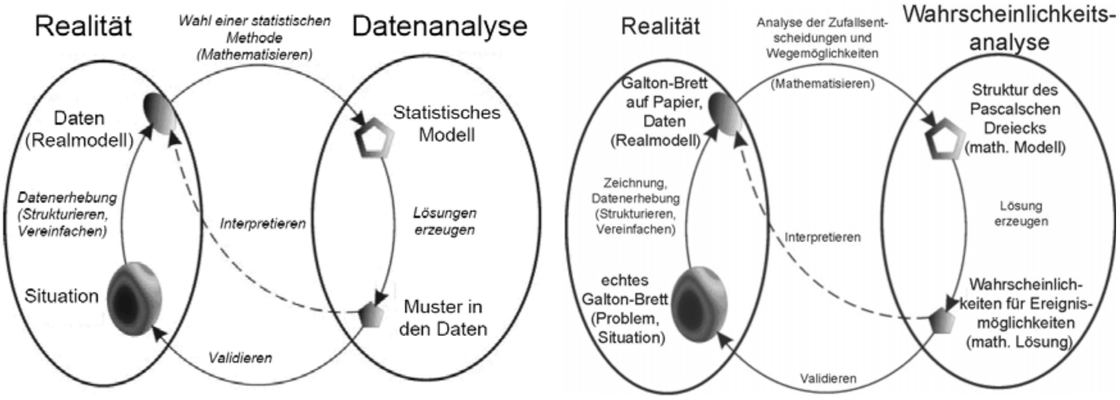
2. Ausblick: Bei vielen statistischen Fragestellungen reicht es aber nicht, den durch Daten repräsentierten Ist-Zustand zu beschreiben. Es wird nach Verallgemeinerungen gefragt, die über den Informationsgehalt der Daten aus der Stichprobe hinausreichen (z. B. Schluss von der Stichprobe einer Volkszählung auf die Eigenschaften des gesamten Volkes). Außerdem sind häufig die Entscheidungen in die Zukunft gerichtet, umfassen also auf konkreten Daten beruhende Prognosen zukünftiger Daten. Sowohl für die Verallgemeinerung wie auch die Prognose benötigt man die Wahrscheinlichkeitsanalyse, um auf der Basis vorhandener Daten (noch) nicht bekannte Daten in ihrer Größe abschätzbar zu machen.

Wie der Realitätsbeschreibung durch Daten liegen auch der Verallgemeinerung und Prognose Modelle von der Realität zu Grunde. Akzeptiert man den Gedanken, dass sich ein wesentlicher Teil der Stochastik auf die Beschreibung aktueller und zukünftiger Daten mit einem realen Kontext richtet (Wild & Pfannkuch, 1999), so besteht die Stochastik, repräsentiert durch Daten- und Wahrscheinlichkeitsanalyse, aus dem Aufstellen, Bearbeiten und Bewerten von Modellen der Realität.

3. Daten- und Wahrscheinlichkeitsanalyse als Modellierungsprozesse

Wird die Daten- und Wahrscheinlichkeitsanalyse aus der Perspektive der Modellbildung betrachtet, so stellt sich die Frage, welche Anforderungen an Modelle der Realität, insbesondere für den Bereich der Schulmathematik, gestellt werden und welche Merkmale sie auszeichnen. Der wesentliche Kern des Modellierens besteht darin, dass nicht die gesamte Komplexität der Ausgangsfragestellung berücksichtigt wird. Es wird bewusst mit Verkürzungen gearbeitet, die nur jeweils relevante Merkmale berücksichtigen. Was aber als relevant angesehen wird, darüber entscheidet das modellbildende Subjekt. Eine eindeutige, zeitlich und inhaltlich allgemeingültige Zuordnung von Original und Modell gibt es daher nicht (*Pragmatisches Merkmal* des allgemeinen Modellbegriffs nach Stachowiak, 1973). „Modelle sind Modelle für jemanden, zu einer bestimmten Zeit und zu einem bestimmten Zweck." (Schupp, 1988, S. 10) Aus diesem Grund sind Modelle nicht hinsichtlich ihrer Richtigkeit, sondern hinsichtlich ihrer Nützlichkeit zu beurteilen (vgl. Box & Draper, 1987). Für die Modellbeurteilung sind nach Stachowiak (1973) zwei weitere Merkmale wesentlich: Die Güte eines Modells bemisst sich zum einen über das *Abbildungsmerkmal* daran, wie gut es die relevanten Eigenschaften des Ausgangssachverhalts wiedergeben kann, und zum anderen über das *Verkürzungsmerkmal* daran, wie gut es der mathematischen Bearbeitung zugänglich ist und intuitive Einsichten zu vermitteln vermag.

In der mathematikdidaktischen Literatur wurde der so genannte mathematische Modellierungskreislauf nunmehr seit einigen Jahrzehnten verschiedentlich vorgestellt und diskutiert. Besondere Aufmerksamkeit erfuhr er insbesondere mit den Veröffentlichungen im Rahmen der PISA-Studien (z.B. Blum et al., 2003). Spezifiziert man den Modellierungskreislauf als stochastischen Modellierungskreislauf ergibt sich – egal, ob allgemein als Datenanalyse oder exemplifiziert am Galton-Brett als Wahrscheinlichkeitsanalyse betrachtet (vgl. Eichler & Vogel, 2009):



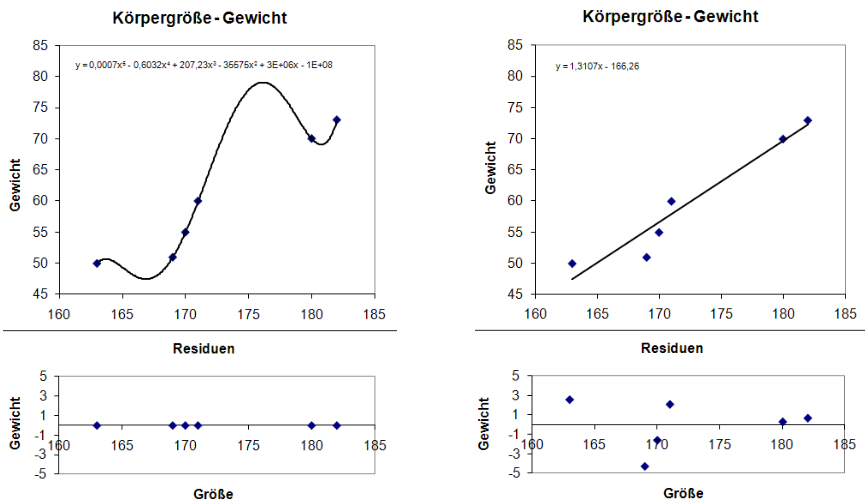
Die Daten bilden die Informationen, die aus Beobachtung, Umfrage oder Experiment hervorgehen, als Kontextzahlen in einem Realmodell der Situation ab. Über Mathematisierungsprozesse des Visualisierens, Simulierens und Formalisierens wird ein mathematisches Modell erstellt, das die Grundlage für statistische Bewertungen oder wahrscheinlichkeitstheoretische Voraussagen bildet. Lösungen, die sich aus Berechnungen auf dieser mathematischen Modellgrundlage ergeben, bilden ein mathematisches Muster ab, das sich im Kontext der Realmodellebene (ggf. unter Hinzunahme neuer Daten) zu bewähren hat. (Hinsichtlich der hier aus Platzgründen notwendigen Verkürzungen sei für ausführlichere Überlegungen auf Eichler & Vogel (2009) sowie Eichler & Vogel (in press) verwiesen.)

Statistische Daten – bereits erhobene wie prognostizierte – hängen in ihrem Entstehungsprozess stets vom Zufall (genauer gesagt: erklärtermaßen als vom Zufall verursacht) ab. Daraus ergibt sich die Variabilität der Daten, die den Kern stochastischen Denkens darstellt. In der Einsicht, dass die Welt nicht streng deterministisch ist, sondern aus einem Mix an Variation und Struktur besteht, gleicht die Datenanalyse einer Suche nach dem Muster in der Variation. Den Zusammenhang zwischen Daten und dem ihnen innewohnenden Paar aus Variabilität und Muster kann man durch die *Grundgleichung der Datenmodellierung* (vgl. z. B. Eichler & Vogel, 2009) ausdrücken, die sich in verschiedener Weise darstellen lässt:

Daten = Muster + Variabilität = Trend + Zufall = Funktion + Residuen

Die additive Struktur der Grundgleichung der Datenmodellierung ist ein pragmatisches Konstrukt, um mit der omnipräsenten Variabilität von Daten fertig zu werden: Der Teil der Datenvariabilität, der erklärt werden kann, wird mit einer deterministischen Komponente zum Ausdruck gebracht. Das, was übrig bleibt, der unerklärte Anteil an Variabilität wird mit dem Konstrukt *Zufall* modelliert. Damit kommt eine stochastische Komponente in die Modellierung der Daten.

Aus der Grundgleichung der Datenmodellierung lässt sich das Ziel der Modellierungsbemühungen beschreiben: Der Musteranteil soll möglichst groß und das, was übrig bleibt, die unerklärte Variabilität möglichst klein sein. Das darf jedoch nicht um den Preis einer sachgerechten Modellierung geschehen, wie folgendes konstruiertes Beispiel (vgl. Vogel & Eichler, 2010) zu einem Zusammenhang von Körpergröße und Gewicht von sechs Erwachsenen zeigt:



Durch die vollständige Datenerfassung verliert das funktionale Modell im linken Beispiel die Möglichkeit, einen sachlich sinnvoll begründbaren Zusammenhang zwischen Körpergröße und Gewicht trendgemäß abzubilden. Dadurch büßt es ein zentrales Merkmal ein, welches ein Modell nach Stachowiak (1973) kennzeichnet: das Abbildungsmerkmal. Dagegen erweckt die rechts dargestellte Regressionsgerade, welche auf der Basis der Minimierung der quadratischen Abweichungen durch den Computer ermittelt wurde, hinsichtlich einer Trendabbildung eine größere Glaubwürdigkeit deshalb, weil sie nichtvorhandene Messwerte in ihrer Größenordnung besser abschätzen lässt (wobei jedoch auch dieses Modell weiter hinterfragt werden kann).

4. Daten und Zufall im Unterricht verknüpfen

Wie sich Daten und Zufall konkret im Unterricht verknüpfen lassen, lässt sich idealtypisch am „Schokolinsen-Beispiel" exemplifizieren (vgl. Eichler & Vogel, 2009). Die Aufgabe besteht für die Schülerinnen und Schüler darin, aus einem empirisch ermittelten Anteil der roten Schokolinsen in einer Tüte eine Prognose für die Grundgesamtheit aller Schokolinsentüten zu diesem Anteil zu erstellen und diese Prognose zu überprüfen. Ein Unterrichtsgang könnte wie folgt aussehen:

1. Datenanalyse: Eine Unterrichtsstunde mit 30 Tüten ergibt, dass durchschnittlich 18 Schokolinsen in einer Tüte sind mit durchschnittlich rund drei Linsen pro Farbe.

2. Modellannahmen: Unter der Annahme, dass eine Tüte hinsichtlich der Farbe unabhängig gefüllt wird, wird geschätzt, dass jede Schokolinse mit der Wahrscheinlichkeit von 1/6 eine bestimmte Farbe hat. Diese Modellannahmen und ihre Grenzen sollen geprüft werden. Auf der Basis dieser Modellannahmen wird die Wahrscheinlichkeit, eine rote Schokolinse aus einer Tüte zu ziehen mit P(rot) = 1/6 (und P(nicht_rot) = 5/6) postuliert. Eine weitere wichtige Modellannahme, die bewusst zu machen ist, ist die Annahme der stochastischen Unabhängigkeit der Ziehungen: Das heißt, die Wahrscheinlichkeiten ändern sich nicht, wenn die nächste Schokolinse aus der Tüte gezogen wird.

3. Simulationsmodell und theoretisches Modell erstellen: Mit dem Computer lässt sich auf der Basis der Modellannahmen ein Simulationsmodell für 10000 Tüten erstellen.

Mithilfe kombinatorischer Überlegungen bzw. mithilfe der (zuvor notwendigerweise behandelten) Binomialverteilung lässt sich ein theoretisches Modell erstellen:

$$\binom{18}{r} \cdot \left(\frac{1}{6}\right)^r \cdot \left(\frac{5}{6}\right)^{18-r}$$

mit r ganzzahlig und $0 \le r \le 18$. Die Gegenüberstellung der Modelle ergibt z.B. folgende Tabellenwerte:

Häufigkeiten bei 10000 Tüten			Theorie Binomialverteilung		
Anzahl	abs.	rel. Häuf.	Anzahl	abs.	Wahrsch.
0	366	0,037	0	376	0,038
1	1369	0,137	1	1352	0,135
2	2329	0,233	2	2299	0,230
3	2405	0,241	3	2452	0,245
4	1871	0,187	4	1839	0,184
5	1031	0,103	5	1030	0,103
6	422	0,042	6	446	0,045
7	155	0,016	7	153	0,015
8	39	0,004	8	42	0,004
> 8	13	0,001	> 8	11	0,001

4. Validierung: Aus den Ergebnissen können Kriterien für die Validierung des zugrunde liegenden theoretischen Modells und/oder des Simulationsmodells in der Realität konstruiert werden. Beispielsweise kann man festlegen, dass das Modell beim Öffnen einer neuen realen Tüte bei 8 oder mehr roten Schokolinsen in einer Tüte abgelehnt und bei weniger als 8 roten Linsen beibehalten wird. In diesem Fall würde man mit einer Wahrscheinlichkeit von weniger als einem Prozent einen Fehler begehen. Die Schülerinnen und Schüler können auch andere Kriterien festgelegen. Sie erfahren dabei, dass dies eine Aushandlungssache ist. Wenn man dabei noch bespricht, welche Arten von Fehlern auftreten können, nämlich die Annahme der Gleichverteilung zu verwerfen, obwohl

diese tatsächlich gegeben ist (Fehler 1. Art) oder an der Annahme der Gleichverteilung festzuhalten, obwohl diese nicht gegeben ist (Fehler 2. Art), so ist man mit diesem Validierungsschritt bei der Grundidee des Hypothesentestens angelangt.

Das Beispiel der Farbverteilung von Schokolinsen zeigt mit den unterschiedlichen Zugangsweisen (Simulation vs. Theorie) in Schritt 3, dass sich Daten- und Wahrscheinlichkeitsanalyse gegenseitig ergänzen. Der Unterschied in beiden Varianten des Modellierens zeigt sich in ihren unterschiedlichen Zugriffspunkten von Empirie und Theorie während dieses Modellierungsschrittes:

- Beim wahrscheinlichkeitstheoretischen Ansatz stehen das theoretische Modell und damit die Wahrscheinlichkeitsanalyse am Anfang. Anschließend erfolgt die Validierung des theoretischen Modells anhand der Analyse empirischer Ergebnisse (etwa von Simulationen), also anhand der Datenanalyse.
- Beim empiriebasierten Ansatz über Simulationen steht die Untersuchung einer empirischen Häufigkeitsverteilung, also die Datenanalyse, am Anfang. Anschließend kann das Ergebnis der Mustersuche in den Daten in ein empirisch begründetes, dennoch aber theoretisches Modell einfließen. Innerhalb dieses Modells erfolgt die Wahrscheinlichkeitsanalyse, dessen Validierung wiederum im Sinne einer Datenanalyse erfolgt.

Wie man zu den gesuchten Wahrscheinlichkeiten oder insgesamt einer Wahrscheinlichkeitsverteilung kommt, hängt von der zugrunde liegenden Problemstellung und der Entscheidung als modellbildende Person ab: Geht man von empirisch ermittelten Häufigkeiten aus oder trägt man (idealerweise begründet) ein theoretisches Modell an die Problemsituation heran – diese Entscheidung ist eine subjektive Entscheidung der Modellbildung.

Die Verknüpfung von Wahrscheinlichkeitsrechnung und Datenanalyse lässt sich in den komplexen Anforderungen des realen Mathematikunterrichts nicht durchgängig verwirklichen. Für die Datenanalyse wie für die Wahrscheinlichkeitsanalyse sind (zum Teil komplexe) mathematisch-technische Werkzeuge notwendig, deren sachgerechte Handhabung als solche erlernt werden muss. Dies darf jedoch nicht zu einer unterrichtlichen Dominanz von ausschließlich kalkülorientierten innermathematischen Aufgaben führen, mit denen der Anspruch eines gehaltvollen Stochastikunterrichts, nämlich dass die Schülerinnen und Schüler lernen, Fragen an empirische Phänomene ihrer erlebten Umwelt zu stellen und mit den zur Verfügung stehenden elementaren mathematischen Mitteln zu beantworten (vgl. oben), nicht eingelöst werden kann. Hans Schupp (1982, S. 210) bringt die Notwendigkeit, Daten- und Wahrscheinlichkeitsanalyse (zumindest immer wieder einmal) auch im Unterricht zu verknüpfen, treffend auf den Punkt: „Statistik ohne Wahrscheinlichkeitsrechnung ist blind […], Wahrscheinlichkeitsrechnung ohne Statistik ist leer […].“

Literatur

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F. & Carstensen, C. H. (2003). In Deutsches PISA-Konsortium (Hrsg.), PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs (S. 47-92). Münster: Waxmann.

Box, G. & Draper, N. (1987). Empirical Model-Building and Response Surfaces. New York: Wiley.

Eichler, A. & Vogel, M. (2009). Leitidee Daten und Zufall – Von konkreten Beispielen zur Didaktik der Stochastik. Wiesbaden: Vieweg+Teubner.

Eichler, A. & Vogel, M. (in press). Daten- und Wahrscheinlichkeitsanalyse als Modellierung. In R. Borromeo Ferri, G. Greefrath & G. Kaiser (Hrsg.), Didaktik des mathematischen Modellierens – Erste Bausteine. Wiesbaden: Vieweg+Teubner

KMK, Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2003). Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. München: Luchterhand.

Schupp, H. (1982). Zum Verhältnis statistischer und wahrscheinlichkeitstheoretischer Komponenten im Stochastik-Unterricht der Sekundarstufe I. Journal für Mathematik-Didaktik, 3 (3/4), 207-226.

Schupp, H. (1988). Anwendungsorientierter Mathematikunterricht in der Sekundarstufe I zwischen Tradition und neuen Impulsen. Der Mathematikunterricht, 34(6), 5-16.

Stachowiak, H. (1973). Allgemeine Modelltheorie. Wien: Springer.

Steinbring, H. & Strässer, R. (Hrsg.) (1981). Rezensionen von Stochastik-Lehrbüchern. Zentralblatt der Mathematikdidaktik (ZDM), 13, 236-286.

Vogel, M. & Eichler (2010). Residuen helfen gut zu modellieren. Stochastik in der Schule, 30(2), 8-13.

Vollrath, H.-J. (1989). Funktionales Denken. Journal für Mathematikdidaktik, 10, 3-37.

Wild, C. & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. International Statistical Review 67(3), S. 223-248.