Von M&Ms und bevorzugten Farben: ein handlungsorientierter Unterrichtsvorschlag zur Leitidee „Daten & Zufall“ in der Sekundarstufe I

JOACHIM ENGEL, HANNOVER & MARKUS VOGEL, LUDWIGSBURG

Zusammenfassung: *Ausgehend von der Frage, ob in Tüten von M&M-Minis bestimmte Farben bevorzugt werden, wird ein handlungsorientierter Unterrichtsvorschlag gemacht, an dem Fragen der Datendarstellung, Analyse und möglicher Verallgemeinerungen auch schon in unteren Klassen der Sekundarstufe I behandelt werden können.*

1. Einführung

Die im Dezember 2003 von der KMK erlassenen „Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss“ haben den Themenbereich „Daten und Zufall“ zu einer von fünf Leitideen des Mathematikunterrichts erklärt. Bildungsplanreformen in einzelnen Ländern wie z.B. in Baden-Württemberg (siehe: www.bildungsstandards-bw.de) stehen der hier zum Ausdruck kommenden inhaltlichen Wertschätzung von Datenanalyse und Wahrscheinlichkeitsrechnung um nichts nach. Gleichzeitig wird von vielen Unterrichtenden ein gravierender Mangel an Unterrichtskonzepten festgestellt, wie die hier geforderten Inhalte des Mathematikunterrichts auszugestalten seien. Im Folgenden wird in Anlehnung an Materialien des US-amerikanischen *Quantitative Literacy Projects* (Burrill et al., 1999) ein Unterrichtsvorschlag gemacht, wie auch schon in unteren Klassenstufen der Sekundarstufe I Schüler Erfahrungen machen können im gesamten Prozess einer empirischen Untersuchung: von der Datenerhebung über die Darstellung und Analyse von Daten bis hin zu Schlussfolgerungen und probabilistisch begründeten Verallgemeinerungen aus Daten.

2. Didaktischer Rahmen

Die aktuelle mathematikdidaktische Diskussion betont den Prozess des Mathematiktreibens als Ziel nachhaltigen Mathematiklernens. Hierbei handelt es sich um prozedurales Wissen, das auf der Grundlage mathematischen Faktenwissens entfaltet werden soll. Bezogen auf den Mathematikunterricht bedeutet dies, dass schulische Lernarrangements nicht vom Ende der Fachsystematik, sondern vom Ausgangspunkt der Mathematik her bestimmt sein sollten. In diesem Zusammenhang ist das Mathematisieren realer Problemstellungen als eine zentrale mathematische Fähigkeit zu nennen. Von wesentlicher unterrichtspraktischer Bedeutung ist dabei, dass die Schülerinnen und Schüler das jeweilige Ausgangsproblem als authentisch, nicht als vom Lehrer „prä-pariert“ erfahren. Im Falle eines nur mittelbaren Problemzugangs besteht die Gefahr, dass die Lernenden versuchen einen noch unbekannten, aber durch die vorweggenommene Problemmodellierung vorgezeichneten Mathematisierungsschritt aufzuspüren, statt diesen Schritt selbst zu tun. In diesem Sinne ist die Forderung Hans Freudenthals zu verstehen, die Phänomenologie mathematischer Begriffe als Ausgangspunkt des Mathematikunterrichts zu sehen: „Unsere mathematischen Begriffe, Strukturen und Vorstellungen sind erfunden worden als Werkzeuge, um die Phänomene der natürlichen, sozialen und geistigen Welt zu ordnen“ (Freudenthal, 1983). So gekennzeichnete Phänomene zeichnen sich durch die Authentizität ihrer Ursprünglichkeit aus.

Phänomenologische Aspekte der Explorativen Datenanalyse

Auf der Basis dieser Überlegungen lässt sich das didaktische Potential der explorativen Datenanalyse entfalten (Borovcnik, 1987). Betrachtet man reale Daten als kontextualisierte Zahlen eines Phänomens im Sinne Freudenthals, besteht das Mathematisieren darin, über die Strukturierung der Daten den phänomenologischen Hintergrund zu entschlüsseln und für eine sachgerechte Interpretation zugänglich zu machen. Diese Interpretation mündet in Annahmen für ein mathematisches Modell der Datenstruktur ein. Bei der Aufnahme und Verarbeitung von empirischer Information erweist sich der Zufallsbegriff als sehr nützlich. Wir lesen einen Trend oder ein Muster aus vorliegenden Daten und erklären alles was übrig bleibt, d.h. die Differenz zwischen Daten und Trend – die Residuen – als Zufall. Zufall in der Analyse empirischer Ergebnisse ist ein Konstrukt, das wir nutzen, um Variabilität in Daten zu modellieren. Das ist der Kern statistischen Denkens. Bei einem von der Statistik herkommenden Ansatz ist der Zufall nicht a priori wie in normativen wahrscheinlichkeitstheoretischen Ansätzen im Modell über eine Wahrscheinlichkeitsverteilung definiert, sondern er muss an die Daten herangetragen werden. Die Gleichung

„Daten = Modell + Zufall“

ist eine Interpretation und ein Konstrukt, das uns Menschen in die Lage setzt, mit der Vielfalt und Variabilität der Daten fertig zu werden und im Nebel der vielen Details das Wesentliche aufzunehmen. Dadurch ist das Zufallsverständnis stark mit der Kompetenz zum statistischen Denken verbunden.

Da wir mit dem hier präsentierten Vorschlag zur propädeutischen Erarbeitung von zentralen Konzepten der Datenanalyse und Stochastik in unteren Klassen der Sekundarstufe beitragen möchten, müssen wir auf formale inferenzstatistische Betrachtungen wie etwa Chi-Quadrat-Tests und komplexere Verteilungsberechnungen verzichten. Dennoch halten wir es für zentral und im Sinne eines Spiralcurriculums auch geboten, Fragen der Verallgemeinerung und der Inferenzstatistik auf einem präformalen Niveau zu behandeln. Mit dem Hilfsmittel der Simulation stehen uns dazu geeignete Werkzeuge zur Verfügung. Auf der Basis eines erstellten Modells werden dabei über Simulationen häufig wiederholt künstliche Daten generiert.

Didaktische Stufenfolge

Auf der Basis der vorangegangenen Überlegungen orientiert sich die nachfolgend beschriebene Unterrichtssequenz an der folgenden, in der Stochastik üblichen, didaktischen Stufenfolge:

Induktion – Mathematisieren:

- Unmittelbarer Datenzugang
- Datenexploration und –strukturierung

Modellannahmen und –synthese Deduktion – Validieren:

- Simulative Überprüfung des Modells
- Modellbewertung und Schlussfolgerungen anhand eines empirischen Wahrscheinlichkeitsbegriffs

Für die unterrichtspraktische Realisierung sei betont, dass die Auflistung keine streng lineare und abgegrenzte Abfolge impliziert, vielmehr greifen verschiedene Stufen im didaktischen Gang jeweils wechselseitig ineinander. Grundsätzlich führen jedoch vorausgehende qualitative Betrachtungen zur Quantifizierung von Schlussfolgerungen, wobei unterschiedliche Problemlösestrategien zum Zuge kommen können.

3. Methodische Analyse

Dem Gedanken der phänomenologischen Anknüpfung und der didaktischen Stufung folgend schlagen wir einen handlungsorientierten Zugang vor, bei der Schülerinnen und Schüler konkrete Erfahrungen mit einem empirischen Untersuchungsprozess machen. Das Beispiel der realen Daten von kleinen farbüberzogenen Schokoladenlinsen verspricht gleichermaßen den Ansprüchen der didaktischen Überlegungen wie auch der Schülermotivation gerecht werden zu können. Dabei kann der im Freudenthal-Zitat genannte Gedanke des Ordnens zu Beginn der Unterrichtseinheit im wahrsten Sinne des Wortes „hand-greiflich“ erfahrbar werden.

Repräsentationsebenen

Daran anknüpfend ist die Methodik des gesamten

Unterrichtsgangs durch fließende Übergänge der enaktiven, ikonischen und symbolischen Repräsentationsebenen gekennzeichnet, um zur Abstraktion der mathematischen Modellierung breite Zugangswege zu eröffnen. Die Modellreflexion wird durch die Möglichkeit der Tabellenkalkulation, ikonische und symbolische Repräsentationen dynamisch zu verknüpfen, unterstützt.

Materialien

Günstig für den unterrichtspraktischen Einsatz sind kleine Tüten von M&M-Minis. Dabei genügen zwei Pakete M&M-Minis, die jeweils ca. 15 kleine Tüten enthalten. Über das kurze Vergnügen des Verzehrens hinaus überrascht auch die unterschiedliche Zusammensetzung und Streuung des Tüteninhalts (z.B. 25-34 Linsen pro Tüte, 0-13 Linsen bei einzelnen Farben). Natürlich lässt sich die Übung auch mit anderen Süßigkeiten oder auch nicht-essbaren Gegenständen durchführen. Unsere Empfehlung für M&M-Minis begründet sich darin,

Abbildung 1: Tüten von M&M-Minis, die für den Unterrichtsvorschlag verwendet wurden.

dass sie sich, wenn nach Farben sortiert, aufgrund ihrer Form in natürlicher Weise zu einem Säulendiagramm sortieren lassen, das einen nahtlosen Übergang zur entsprechenden ikonischen Repräsentation gewährleisten. Außerdem treten M&M-Minis in sechs verschiedenen Farben auf. Diese Tatsache eignet sich zu einem simulativen Nachahmen eines erstellten Modells. Legen die Daten die Hypothese einer Gleichverteilung nahe, so lassen sich dann mittels (Farb-)Würfel problemlos viele Daten simulieren, die jeweils mit den ursprünglichen Daten verglichen werden können. Die Repräsentationsmöglichkeiten eines interaktiven Arbeitsblatts der Tabellenkalkulation bieten nahe liegende grafische Anknüpfungspunkte an das Datenmaterial der M&M-Minis.

Unterrichtsgang

Jeder Schüler und jede Schülerin erhält zunächst eine Tüte („die Daten“) für eine eigene Untersuchung: Vor dem Öffnen wird zunächst eine Vermutung über die mögliche Zusammensetzung der geschlossenen Tüte formuliert und festgehalten. Diese Vermutung wird anschließend anhand der eigenen Tüte überprüft. Vermutung und Ergebnis werden einander in Tabelle und Diagramm gegenübergestellt. Zusätzlich bietet sich an, dass die Schülerinnen und Schüler das Ergebnis ihrer eigenen Untersuchung schriftlich formulieren. Im weiteren Verlauf werden in der gleichen Vorgehensweise die Daten erst für Vierergruppen und schließlich für die gesamte Klasse zusammengerechnet. Dabei werden möglicherweise aufgrund der ursprünglichen Stichprobe (vom Umfang ca. 30) vollzogene Schlüsse und Interpretationen aufgrund der aggregierten Stichproben, die aus den Daten der Vierergruppen beziehungsweise aller Schülerinnen und Schüler besteht, revidiert. Schüler sollen dabei auch über eine ihnen typisch erscheinende Zusammensetzung einer Tüte von M&M-Minis reflektieren und erklären, wie sie zu ihrer Wahl einer typischen Tüte gekommen sind, und wie sie Abweichungen von der „typischen Tüte“ messen. Die Schüler müssen dabei daran erinnert werden, die Süßigkeiten nicht zu essen, bevor alle Daten aufgezeichnet worden sind, weil sonst die Ergebnisse verzerrt sind. Auch sollte sichergestellt werden, dass Schüler ihre Farbvorhersagen machen, *bevor* sie die Tüten öffnen.

Nachdem die Schülerinnen und Schüler die Zusammensetzung einer typischen Tüte beschrieben, begründet und die Relation ihrer eigenen Tüte zur typischen Tüte quantifiziert und dargestellt haben, formulieren sie eine Vermutung über die Produktionsparameter der Herstellung bezüglich der Farbverteilung. Im Falle der Gleichverteilungsannahme kann dieser Vermutung im nachfolgend angegebenen Excel-Arbeitsblatt anhand von mehreren Simulationen nachgegangen werden. Eine größere Zahl von Simulationen bildet die Grundlage für die quantitative Überprüfung der Schlussfolgerungen zum Modell einer Farbverteilung.

Zur Leistungsdifferenzierung kann auf der Motivation einer Farbpräferenz (vgl. zum Beispiel rote Gummibärchen) aufbauend als Vertiefungsmöglichkeit der Frage nachgegangen werden, wie viele Tüten M&M-Minis nötig sind, um mit einer gewissen Zuversicht eine bestimmte Anzahl einer Farbe zu erhalten. Die bewusst vage und präformal gewählten Formulierungen sollen zum Nachdenken einladen, das erst allmählich zu einer Präzisierung der Quantifizierung von Unsicherheit hinführt.

4. Fragenkatalog

Der untenstehende Fragenkatalog (**siehe nächste Seite**) kann den inhaltlichen Unterrichtsgang nachzeichnen. Er sollte nicht als „Unterrichtsfahrplan“ missgedeutet werden, die konkreten didaktisch-methodischen Entscheidungen bleiben der Lehrkraft überlassen, um eine optimale unterrichtspraktische Passung zu gewährleisten.

5. Antwortmöglichkeiten

Die Antworten zu Frage 2 werden unterschiedlich ausfallen, da jeder Schüler seine eigenen Daten hat, die sich von den Daten der Mitschüler deutlich unterscheiden können. Erfahrungsgemäß ist bei einer vorliegenden Stichprobe von ca. 30 M&M-Minis mit sechs unterschiedlichen Merkmalsausprägungen die Variabilität recht groß. Die relativen Häufigkeiten der jeweiligen Anzahlen stabilisieren sich aber bei einer Klassenstärke von 20 oder mehr Schülern merklich. Die Frage nach dem Zustandekommen der beobachteten Unterschiede in den Farbverteilungen ist bewusst offen. In einem datenorientierten Zugang zur Stochastik wird Zufall nicht a priori formal durch Einführung einer Zufallsvariable in einem (Standard-) Modell definiert, sondern dient als Annahme um Variabilität in Daten zu modellieren. Der Zufallsbegriff wird also als Konstrukt herangezogen, um die in allen empirischen Daten omnipräsente Variabilität handhaben zu können. Im hier betrachteten Fall besteht eine nahe liegende Vorstellung in der Annahme, dass der Hersteller sehr viele Schokolinsen nach fest vorgegebenen Proportionen der Farben produziert, diese in einem großen Behälter zwischenlagert, von wo aus ca. 30 Exemplare zufällig ausgewählt und in kleine Tüten abgefüllt werden. Auch die Antworten zu Frage 3c können recht unterschiedlich ausfallen. Für die individuellen Daten (pro Schüler) liegt eine Darstellung der Daten als Säulendiagramm sehr nahe. Bei Frage 4 werden die Antworten ebenfalls wohl recht unterschiedlich ausfallen. Diese Frage, ein Zwischenschritt zur Modellbildung, bereitet schon implizit Frage 6 vor. Daher sollten hier die Antworten eine Begründung einschließen, warum gerade die vorgeschlagene Wahl als typisch angesehen wird. In der Besprechung werden dann sicherlich auch die arithmetischen Mittel der individuellen Durchschnitte für jede Farbe von M&M angesprochen werden. Mögliche Vorschläge umfassen ganzzahlig gerundete Mittelwerte, Modalwerte oder Mediane. Werden bei Frage 5a absolute Abweichungen notiert, mögen Schüler diese Differenzen einfach aufaddieren. Eine andere Möglichkeit ist das arithmetische Mittel der Abweichungen. Vielleicht ersinnen Schüler auch noch ganz andere Methoden, die akzeptabel sind. Falls negative Zahlen als Abweichungen notiert werden, dann führt eine Addition dieser Zahlen in die Irre. Denn diese Summe kann 0 sein, auch wenn die Abweichungen tatsächlich recht groß sind. Bei 5b) umfassen mögliche Antworten Säulendiagramme, wobei der Unterschied als gestapeltes Säulendiagramm dargestellt ist und Pfeile andeuten, ob die Differenz positiv oder negativ ist. Alternativ ist auch ein gepaartes Säulendiagramm möglich, das zu jeder Farbe nebeneinander stehende Säulen anzeigt.

Wie viele Schokolinsen von jeder Farbe erwartest Du in einer Tüte M&M-Minis?

1.) Bevor Du die Tüte öffnest, fülle die Spalte „Geschätzte Anzahl“ in der Tabelle aus, indem Du rätst, wie viele M&Ms sich insgesamt und von jeder der sechs einzelnen Farben in Deiner Tüte befinden.

Daten zu Farben von M&Ms

Farbe	Geschätzte Anzahl	Tatsächliche Anzahl	Mittelwert in der Klasse	Differenz (Tatsächliche Anzahl minus Mittelwert der Klasse)
Grün				
Rot				
Orange				
Blau				
Gelb				
Braun				
Insgesamt				

2.) Datenpräsentation

a. Sortiere und zähle die Anzahl der M&Ms für jede Farbe und notiere die Werte in der Spalte „Tatsächliche Anzahl“.
b. Mach deine Ergebnisse deinen Mitschülern verfügbar. Stellst Du Unterschiede zu den Ergebnissen Deiner Mitschüler fest? Wie können diese Unterschiede zustande gekommen sein?

3.) *Datenanalyse*

a. Ermittle die durchschnittliche Anzahl von jeder Farbe und trage sie in die Spalte „Mittelwert in der Klasse“ ein.
b. Errechne die Differenz zwischen tatsächlicher Anzahl und Klassenmittelwert und notiere die Werte in der letzten Spalte der Tabelle.
c. Wie könnte man die Ergebnisse geeignet graphisch darstellen?

Schlussfolgerungen

4.) Beschreibe und begründe in ein paar Sätzen, wie viele M&Ms von jeder Farbe Du in einer „typischen“ Tüte erwarten würdest.

5.) Errechne, wie nahe deine Anzahlen zu den Anzahlen der typischen Tüte sind.

a. Nutze die letzte Spalte der Tabelle und erdenke dir eine Formel oder eine Zahl, die die Nähe deiner Werte zu den Werten der typischen Tüte ausdrückt.

b. Erstelle eine graphische Darstellung für die Nähe deiner Farbanzahlen zum Klassendurchschnitt

6.) Werden vom Hersteller bestimmte Farben bevorzugt, oder kommen alle Farben gleich oft vor? Begründe Deine Vermutung.

Erweiterungen

7.) Wie viele kleine Päckchen musst du öffnen, um mindestens 25 grüne M&Ms zu erhalten? Ergibt diese Anzahl der Päckchen jedes Mal exakt 25 grüne M&Ms? Erkläre!

Frage 6 führt zur Modellbildung. Was hat sich der Hersteller gedacht? Werden von der Firma bestimmte Farben bevorzugt, oder werden M&M-Minis in gleichem Umfang in allen sechs Farben hergestellt? Diese Frage führt direkt zur Verallgemeinerung von der Stichprobe der vorliegenden M&M-Minis zur Grundgesamtheit ALLER produzierten M&M-Minis. Hier kann in den unteren und mittleren Klassenstufen keine formale inferenzstatistische Begründung geliefert werden. Dennoch lassen sich zumindest auf einer qualitativen Ebene Positionen und Beurteilungen begründen. Am Anfang steht der Schritt von den Daten zum Modell, das formuliert werden kann etwa als: „*Die Firma bevorzugt keine Farben, sondern stellt M&M-Minis in gleichen Häufigkeiten her*." Eventuell verleiten die Daten auch zu einer anderen Modellierung, die dann vielleicht formuliert werden als „*Jeweils 20% aller produzierten M&M-Minis sind braun, gelb, orange und rot, und jeweils 10% sind blau und grün*."

… und vom Modell per Simulation zu den Daten

Wie lässt sich dieses Modell überprüfen? Eine Möglichkeit, die nicht auf formal-mathematischer Herleitung basiert, besteht im Erstellen einer großen Zahl von (mindestens 100) Simulationen. Der mathematische Hintergrund hierfür ist das Gesetz der großen Zahlen. Dabei mag es sinnvoll sein, zunächst einen physischen Zufallszahlengenerator wie z.B. einen Würfel einzusetzen, dessen Wirkweise und Zufallscharakter den Schülern vertraut ist. Im Fall der Annahme einer Gleichverteilung kann hier zunächst mit einem (Farben-)Würfel gearbeitet werden. Eine Tüte M&M-Minis wird dann durch 30maliges Werfen des Würfels simuliert. Da hier jedoch Simulationen in größerem Umfang notwendig sind, um mit dem Gesetz der großen Zahlen argumentieren zu können, ist der Einsatz von Software ratsam. Wir haben dazu ein Excel-Arbeitsblatt erstellt, mit dem bei angenommener Gleichverteilung die Verteilung der Farben von 30 M&M-Minis (individuelle Tüte) und 25 mal 30 = 750 M&M-Minis (Daten für die gesamte Klasse) erzeugt werden. Dabei gehen wir von einer festen Anzahl von 30 „M&M-Minis" pro Tüte aus und ignorieren die Tatsache, dass in den tatsächlichen Tüten die Anzahl der M&M-Minis selbst schwankt. Wiederholte Simulationen, mit ihren Ergebnissen als Säulendiagramm repräsentiert, vermitteln einen Eindruck von der zufallsbedingten Variabilität, mit der die sechs Farben bei einer Gleichverteilungsannahme verteilt sein können. Diese Variabilität ist bei den einzelnen Tüten erheblich größer als bei den aggregierten Daten für die ganze Klasse. Bei 750 Daten sind die Prozentsätze, die auf die einzelnen Farben fallen, viel näher bei einander als bei nur 30 Daten. Eine mathematisch präzisere Fassung dieser Feststellung (die aber kaum mehr Sekundarstufe 1 tauglich ist) besagt, dass die Standardabweichung bei Erhöhung des Stichprobenumfangs um das 25-fache – nämlich von 30 auf 750 – um das 5-fache reduziert wird. Abbildung 2 zeigt Säulendiagramme für relative Häufigkeiten der einzelnen Farben bei einer mit Excel durchgeführten Simulation mit Stichprobenumfang 30 (dunkel) bzw. 750 (hell).

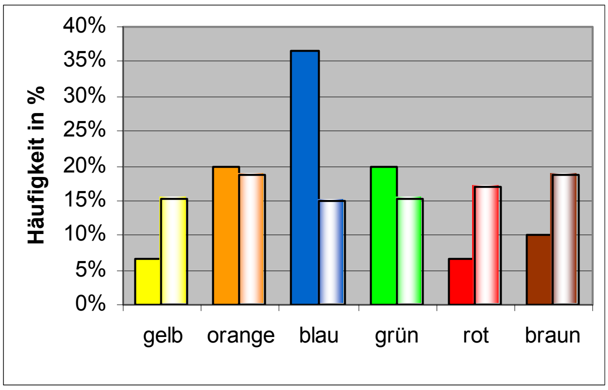

Abbildung 2: Ergebnis einer Simulation: Die Säulenhöhe gibt die relative Häufigkeit an, wie oft die jeweilige Farbe bei 30-facher (dunkel) bzw. 750-facher (hell) gewählt wurde.

Interessant und weiterführend ist auch die Diskussion der Frage 7. Naive Antworten mögen lauten, dass – Gleichverteilung der Farben vorausgesetzt – 5 bis 6 Tüten genügen sollten, um 25 grüne M&M-Minis zu erhalten. Tatsächlich kann man sich aber niemals sicher sein, dass auch nach 100 Tüten 25 grüne M&M-Minis erhalten wurden. Man könnte hier die Schüler ihre qualitativen Urteile einfach diskutieren lassen – *„wenn mir 25 grüne M&M-Minis wichtig sind, dann kaufe ich 8 oder 9 Tüten"* – eine Quantifizierung kann hier wiederum über eine größere Zahl von Simulationen erfolgen, und stellt eine Nutzung (eventuell auch Einführung) des empirischen Wahrscheinlichkeitsbegriffs dar. Dazu werden 30 Zufallszahlen zwischen 1 und 6 erzeugt (ohne Einschränkung identifizieren wir die grüne Farbe mit der 6), die Anzahl der 6-en gezählt, und sooft wieder 30 Zufallszahlen erzeugt, bis die Anzahl der 6-en die Zahl 25 erreicht oder überschritten hat. Wir notieren die Anzahl der dazu notwendigen Wiederholungen, bis 25 grüne M&M-Minis erhalten wurden. Dieses Experiment wird z.B. 100 Mal wiederholt. Die somit erhaltenen 100 Werte stellen wir im Histogramm dar und berechnen das arithmetische Mittel als Annäherung für die – unter der Modellannahme der Gleichverteilung – zu erwartende Anzahl zu öffnender Tüten M&M-Minis. Das Histogramm vermittelt einen qualitativen Eindruck der hier zu beobachtenden Zufallsschwankungen, der mittels entsprechender Streuungsparameter quantifiziert werden kann (siehe Abbildung 3). Ebenso können empirische Wahrscheinlichkeiten errechnet werden, mit welcher Wahrscheinlichkeit eine bestimmte Zahl von Tüten ausreicht, wenn man z.B. 25 grüne M&M-Minis haben möchte. Im Beispiel der in Abbildung 3 dargestellten simulativ erhaltenen Werte gilt z.B.:

P(6 Tüten reichen aus) $\approx$ 53/60 $\approx$ 88,3 %,

P(7 Tüten reichen aus) $\approx$ 59/60 $\approx$ 98,3 %.

Die beiden angegebenen Excel-Arbeitsblätter dienen nur als Arbeitsvorschläge. Für den unterrichtlichen Einsatz können diese situativ und den Vorkenntnissen der Schülerinnen und Schüler entsprechend angepasst werden.

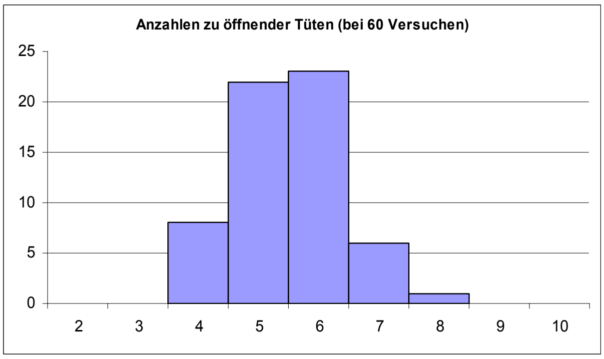

Abbildung 3: Histogramm der simulativ erhaltenen Werte zur Frage: Wie viele Tüten mit 30 M&M-Minis muss ich kaufen, wenn ich 25 grüne M&M-Minis haben möchte (bei Annahme des Gleichverteilungsmodells).

6. Ergänzungen

Modellvalidierung

Eine Internet-Recherche ergibt, dass entsprechend der M&M-Produktphilosophie für M&M-Minis das Gleichverteilungsmodell den Vorgaben des Herstellers entspricht. Bei anderen ähnlichen Produkten desselben Herstellers sind andere, von der Gleichverteilung abweichende Modellannahmen angemessener, siehe:
http://ch.mms.com/mms/swi_ger/about/products/crispy/default.htm

Formalisierung

Wir haben oben bewusst auf eine mathematische Formalisierung verzichtet, weil wir davon überzeugt sind, dass eine zu frühe Einführung komplexerer Formeln für die Entwicklung tragfähiger Grundvorstellungen kontraproduktiv ist. Die neuen curricularen Vorgaben bieten wichtige Chancen, Denk- und Argumentationsweisen der Inferenzstatistik auch schon in unteren und mittleren Klassenstufen einzuführen. Dabei muss aber – im Sinne eines Spiralcurriculums – altersgemäß vorgegangen werden. Eine Konzentration auf komplexere algebraische Ausdrücke halten wir hingegen für die anvisierte Zielgruppe von Schülerinnen und Schülern zwischen der 6. und 9. Klasse für kontraproduktiv. In diesem Abschnitt geben wir (für das Hintergrundwissen des Lehrers oder für eine Behandlung der hier aufgeworfenen Fragenstellungen in höheren Klassenstufen) Formeln für eine analytische Bearbeitung. Unter Annahmen der stochastischen Unabhängigkeit der Farben ist hier eine Multinomialverteilung $M(n, p_1, \ldots, p_k)$ das angemessene Modell. Im hier interessierenden Fall von $k = 6$ Farben, die alle mit derselben Wahrscheinlichkeit $p = \frac{1}{6}$ auftreten, ergibt sich für den Zufallsvektor $(X_1, \ldots, X_6)$ der Anzahlen des Auftretens der sechs Farben bei $n = 30$ Schokolinsen

$$P(X_1 = x_1, \ldots, X_6 = x_6) = \frac{30!}{x_1! \cdot \ldots \cdot x_6!} \frac{1}{6^{30}}.$$

Schauen wir uns an, wie die Farbverteilung einer speziellen Farbe – z.B. die Anzahl der blauen Schokolinsen – variiert, so reduziert sich das Modell auf eine Binomialverteilung mit den Parametern $n = 30$ und $p = \frac{1}{6}$.

Die einzelnen Komponenten der Multinomialverteilung sind nicht stochastisch unabhängig, sondern es gilt für die Kovarianzen ($i \neq j$)

$Cov(X_i, X_j) = -np_i p_j$ (siehe z.B. Hinderer, 1972).

Im hier vorliegenden Fall mit $n = 30$, $p = \frac{1}{6}$ ergibt sich $Cov(X_i, X_j) = -\frac{5}{6}$ und somit gilt für den Korrelationskoeffizient $\rho(X_i, X_j) = -\frac{1}{5}$.

Die negative Korrelation überrascht nicht: Da die Gesamtzahl der Schokolinsen auf 30 festgesetzt ist, so wird tendenziell die Zahl der blauen Linsen niedrig sein, wenn die Zahl der roten Linsen hoch ist. Folgende Tabelle mit den (kumulativen) Wahrscheinlichkeiten gibt Auskunft über das Auftreten der jeweiligen Anzahlen:

k	$B(30, \frac{1}{6}; k)$	$\sum_{i \le k} B(30, \frac{1}{6}; i)$
0	**0,0042127**	0,0042127
1	**0.02527632**	0.02948904
2	**0.07330133**	0.1027904
3	**0.1368292**	0.2396195
4	**0.1847194**	0.4243389
5	**0.1921081**	0.616447
6	**0.1600901**	0.7765371
7	**0.1097761**	0.8863132
8	**0.06312124**	0.9494344
9	**0.03085927**	0.9802937
10	**0.0129609**	0.9932546
..		

Aus dieser Tabelle ist z.B. zu entnehmen, dass unter den obigen Modellannahmen in ca. 10% der Tüten mit höchstens 2 blauen M&M-Minis zu rechnen ist und in ca. 2% (1-98%) der Tüten mit mindestens 9 blauen Schokolinsen.

Möglicherweise interessiert aber weniger die Verteilung einer speziellen Farbe (wie hier der Farbe blau). Vielmehr erweckt das sehr niedere Vorkommen *irgendeiner* Farbe bzw. das häufige Vorkommen *irgendeiner* Farbe unsere Aufmerksamkeit. Wir haben daher unter der Multinomialverteilungsannahme die Verteilung der minimalen Anzahl von Schokolinsen errechnet mittels:

$$P(Min = k) = P(Min \geq k) - P(Min \geq k+1),$$

wobei

$$P(Min \geq k) = \sum_{x_1 \geq k,\ldots,x_6 \geq k} P(X_1 = x_1,\ldots,X_6 = x_6)$$

Mit Hilfe von MAPLE führte das zu folgendem Resultat:

Minimum	Wahrscheinlichkeit
0	0,02520
1	0,14625
2	0,35827
3	0,37422
4	0,09566
5	0,00040

Entsprechende Berechnungen haben wir für die Verteilung der maximalen Anzahl von Schokolinsen durchgeführt.

$$P(Max = g) = P(Max \leq g) - P(Max \leq g-1),$$

wobei

$$P(Max \leq g) = \sum_{x_1 \leq g,\ldots,x_6 \leq g} P(X_1 = x_1,\ldots,X_6 = x_6)$$

Maximum	Wahrscheinlichkeit
5	0,00040
6	0,08803
7	0,31015
8	0,30722
9	0,17660
10	0,07715
≥ 11	0,04045

Bei 30 Schokolinsen in sechs verschiedenen Farben sind die beiden Ereignisse „Minimale Anzahl einer Farbe =5“ äquivalent zum Ereignis „maximale Anzahl einer Farbe=5“, weil in diesem Fall alle Farben genau 5 mal vertreten sein müssen. Unter einer Multinomialverteilung geschieht dies mit einer Wahrscheinlichkeit von $p = 0{,}00040$.

Bei allen bisherigen Überlegungen sind wir von einer festen Anzahl von $n = 30$ Schokolinsen pro Tüte ausgegangen. Diese Zahl ist aber nicht fix. Da diese Zahl variiert, ist die obige Verteilung lediglich eine bedingte Wahrscheinlichkeitsverteilung, gegeben $N = 30$, wobei jetzt die Zahl der Schokolinsen pro Tüte als Zufallsgröße N dargestellt ist. Wie können wir die Anzahl der Linsen pro Tüte modellieren? Eine Poissonverteilung scheidet als Modell aus, da bei dieser Verteilung Erwartungswert und Varianz identisch sind. Eine Untersuchung von 103 Tüten M&M-Minis ergab aber einen empirischen Mittelwert von 29,8 Schokolinsen mit einer empirischen Varianz von 2,038. Eine Poissonverteilung streut also viel mehr als die vorliegenden Daten. Geeigneter ist ein Normalverteilungsmodell N(29,8; 2,038) oder – da die Daten ja diskrete natürliche Zahlen sind – ein Binomialmodell B(n, p). Die Parameter n und p eines geeigneten Binomialmodells errechnen sich dabei aus $np(1-p) = 2{,}038$, $np = 29{,}8$. Hieraus folgt $p = 0{,}93$ und schließlich $n = 32$. Abbildung 4 zeigt ein Histogramm der Daten mitsamt eingepasster Normalverteilung, erstellt mit FATHOM.

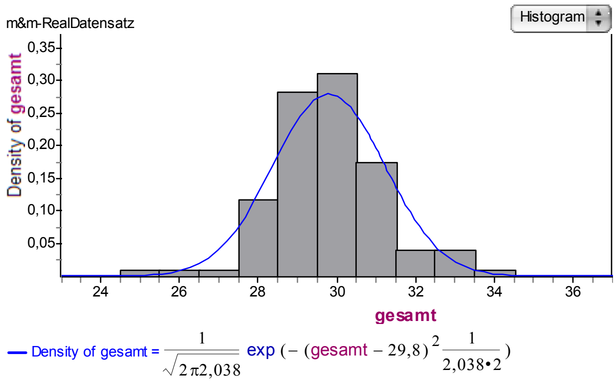

Abbildung 4: Histogramm der Anzahl der Schokolinsen in 103 Tüten M&M-Minis einschließlich eingepasster Normalverteilungskurve.

Das Normalverteilungsmodell passt zumindest als plausible Annäherung gut, da die Daten in etwa symmetrisch sind und die Wölbung (Kurtosis) in den Daten in etwa der Normalverteilung entspricht.

Die Verteilung des Zufallsvektors (N, X_1, …, X_6) ergibt sich jetzt aus dem Produkt von Randverteilung von N und bedingter Verteilung von (X_1, …, X_6) gegeben der Wert von N, d.h.

$$P(N = n, X_1 = x_1,\ldots,X_k = x_k)$$

$$= \frac{1}{\sqrt{2\pi \cdot 2{,}038}} e^{-\frac{(n-29{,}9)^2}{2 \cdot 2{,}038}} \frac{n!}{x_1! \cdot \ldots \cdot x_6!} \frac{1}{6^n}$$

Das Warten auf z.B. 25 grüne Schokolinsen ist ein Paradebeispiel für eine Modellierung mittels negativer Binomialverteilung NB(r, p). Bei unabhängigen Wiederholungen eines Experiments gibt die negative Binomialverteilung die Wahrscheinlichkeiten für die Wartezeiten bis zum r-ten Erfolg an, wenn sich pro Durchgang ein Erfolg mit Wahrscheinlichkeit p ergibt. Es ist

$$P(Y = k) = \binom{k+r-1}{k} p^r (1-p)^{k-r}.$$

Für den Erwartungswert gilt $E(Y) = \frac{r}{p}$,

Wobei die Zufallsvariable Y die Wartezeit bis zum r-ten Erfolg angibt. Im vorliegenden Fall errechnet sich mit $r = 25$ und $p = 1/6$ ein Erwartungswert von 150. Allerdings ist hier zu beachten, dass die Schokolinsen nicht einzeln sequentiell „aufgedeckt" werden, sondern komplette Tüten mit ca. 30 Schokolinsen werden auf einmal geöffnet. Wie groß ist die Wahrscheinlichkeit, mit t Tüten 25 grüne Schokolinsen zu haben? Diese Wahrscheinlichkeit errechnet sich kumulativ mittels $P(Z \geq 25)$, wobei Z eine binomialverteilte Zufallsvariable $Z \sim B\left(30t, \frac{1}{6}\right)$ ist.

Zum Abschluss sei aber noch einmal betont, dass die Überlegungen dieses Abschnitts der theoretischen Auseinandersetzung vor allem des Lehrers dienen, und dass die zum Teil aufwändigen Formeln nicht vom Einsatz des hier vorgeschlagenen Unterrichts schon ab Klasse 6 abschrecken sollen.

Software

Die beiden Excel-Arbeitsblätter sind verfügbar unter dem Namen

`m&m_simulation1.xls`

`m&m_simulation2.xls`

und können heruntergeladen werden von der URL:

`www.ph-ludwigsburg.de/mathematik/personal/vogel/excel`

7. Literatur:

Borovcnik, M. (1987). Zur Rolle der Beschreibenden Statistik. Teil II. *Mathematica didactica* 10, S. 101-117

Burrill, J. et al. (1999). Mathematics in a World of Data. Data-Driven Mathematics. Dale Seymour Publications, White Plains, New York

Freudenthal, Hans (1983). Didactical phenomenology of mathematical structures. Dordrecht: Reidel

Hinderer, K (1972): Grundbegriffe der Wahrscheinlichkeitstheorie. Springer: Heidelberg

Joachim Engel
Fakultät für Mathematik und Physik
Universität Hannover
Welfengarten 1
30167 Hannover
engel@math.uni-hannover.de

Markus Vogel
Pädagogische Hochschule Ludwigsburg
Reuteallee 46
71634 Ludwigsburg
vogel_markus@ph-ludwigsburg.de