Abrücken von der Gleichverteilung bei Fächerbesetzungen: Stochastik trifft Analysis

NORBERT HENZE, KARLSRUHE, UND REIMUND VEHLING, HANNOVER

Zusammenfassung: *Dieser Aufsatz behandelt eine bekannte Problemstellung, die in einer speziellen Einkleidung als Geburtstagsproblem bekannt ist. In n vorgegebene Fächer fallen zeitlich sukzessiv und unabhängig voneinander Teilchen. Die Zufallsgröße X gebe die Anzahl der Teilchen an, die nötig sind, um erstmals eine Doppelbesetzung irgendeines Faches zu erreichen. Es geht dann um die Berechnung der Wahrscheinlichkeiten $\mathbb{P}(X = k)$, $k \in \{2, \ldots, n+1\}$. Standardmäßig wird hierbei angenommen, dass jedes Teilchen mit gleicher Wahrscheinlichkeit in jedes der Fächer fällt. Wir widmen uns der spannenden Frage, welche Konsequenzen ein Abrücken von dieser Gleichverteilungsannahme hat.*

1 Einleitung

Das wohl zum ersten Mal von Richard von Mises (1939) publizierte sog. *Geburtstagsproblem* erfreut sich ungebrochener Attraktivität, siehe z.B. Arnold und Glass (2013), Barth und Haller (2012), Barth und Haller (2013), Gorenflo und Sauer (2018), Riehl (2006), Riehl (2014) und Schrage (1990). Es gilt als Musterbeispiel dafür, wie unsere Intuition bei der Beurteilung von Wahrscheinlichkeiten versagen kann. Üblicherweise blendet man dabei Schaltjahre aus und macht die Annahme, dass alle Personen unabhängig voneinander mit gleicher Wahrscheinlichkeit an jedem der 365 Tage des Jahres Geburtstag haben. Wenn man von dieser Gleichverteilungsannahme abweicht, so wird die Wahrscheinlichkeit, dass von k Personen mindestens zwei am gleichen Tag Geburtstag haben, für jedes $k \in \{2, \ldots, 365\}$ größer (s. z.B. Henze (2015), Abschn. 3). Abstrahiert man dieses Problem, so liegen allgemein n Fächer vor, in die sukzessive und unabhängig voneinander zufällig Teilchen fallen.

Im Fall $n = 365$ entsprechen die Fächer den Tagen des Jahres, und die Teilchen sind Personen. Eine Kollision bedeutet hier, dass ein Mehrfachgeburtstag auftritt. Im Fall $n = \binom{49}{6}$ stehen die Fächer für die lexikographisch durchnummerierten möglichen Gewinnzahlen-Kombinationen beim Lotto 6 aus 49, und da tritt eine Kollision auf, wenn erstmals sechs Gewinnzahlen gezogen werden, die schon früher einmal auftraten (s. Henze (1995)). In der Kryptologie sind die Fächer Wertebereiche sog. *Hashfunktionen*, und die Teilchen sind Dokumente, die mithilfe einer Hashfunktion den sog. *Hash-Adressen* zugeordnet werden (s. z.B. Knuth (1998), S. 513 ff.). Eine Kollision tritt hier auf, wenn zwei verschiedene Dokumente auf den gleichen Hashwert abgebildet werden.

Zentraler Gegenstand dieses Aufsatzes ist eine mit X bezeichnete Zufallsgröße. Diese modelliert die Anzahl der Teilchen, die nötig sind, bis erstmals ein Teilchen in ein Fach fällt, das schon besetzt ist, also eine *Kollision* stattfindet.

Üblicherweise wird bei Kollisionen im Fächer-Modell eine Gleichverteilung unterstellt. Diese Annahme bedeutet, dass jedes Teichen mit gleicher Wahrscheinlichkeit $\frac{1}{n}$ in jedes der n Fächer fällt. Schon mit dieser Standard-Modellierung kann man auf allen Anforderungsniveaus sehr viel entdecken. Eine selbstdifferenzierende Aufgabe ergibt z. B. die folgende Problemstellung:

a) Ermitteln Sie für einige Werte von n jeweils die Wahrscheinlichkeitsverteilung von X sowie den Erwartungswert von X.

b) Für den allgemeinen Fall ergibt sich

$$\mathbb{P}(X = k) = \binom{n}{k-1} \cdot (k-1)! \cdot (k-1) \cdot \frac{1}{n^k}$$

für jedes $k = 2, 3, \ldots, n+1$. Begründen Sie diese Formel.

Hier sollte die Mehrzahl der Schülerinnen und Schüler zumindest den Fall $n = 2$ erfolgreich bearbeiten können. Wichtig ist dabei die Erkenntnis, dass die Zufallsgröße X im Fall $n = 2$ nur die Werte 2 und 3 annehmen kann, denn spätestens im dritten Versuch ergibt sich eine Kollision – eine schöne Anwendung des Schubfachprinzips. Sicherlich wird zur Lösung auch ein Baumdiagramm verwendet, s. Abb. 1. Damit lässt sich erkennen, dass das zweite Teilchen in das gleiche Fach wie das erste Teilchen fallen muss. Mithilfe der beiden Pfadregeln ergibt sich

$$\mathbb{P}(X = 2) = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}.$$

Mit $\mathbb{P}(X = 3) = 1 - \mathbb{P}(X = 2)$ folgt $\mathbb{P}(X = 3) = \frac{1}{2}$ und $\mathbb{E}(X) = 2.5$.

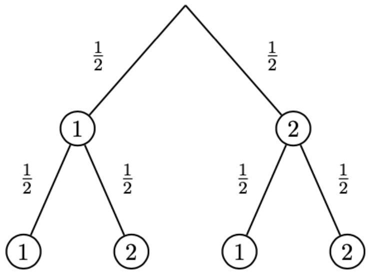

Abb. 1: Baumdiagramm für den Fall $n = 2$

Schreiben wir (i, j), wenn das erste Teilchen in Fach i und das zweite in Fach j fällt, so tritt im Fall $n = 3$ die erste Kollision nach nur zwei Teilchen auf, wenn sich eines der drei Ergebnisse $(1,1)$, $(2,2)$ oder $(3,3)$ einstellt. Nach den Pfadregeln ist die Wahrscheinlichkeit hierfür gleich $3 \cdot \left(\frac{1}{3}\right)^2$, und somit gilt $\mathbb{P}(X = 2) = \frac{1}{3}$. Damit im Fall $n = 3$ die erste Kollision erst mit dem vierten Teilchen erfolgt und somit das Ereignis $\{X = 4\}$ eintritt, müssen die drei ersten Teilchen in verschiedene Fächer fallen. Notieren wir (i, j, k), wenn das erste Teilchen in Fach i, das zweite in Fach j und das dritte in Fach k fällt, so tritt das Ereignis $\{X = 4\}$ genau dann ein, wenn eines der Ergebnisse $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$ oder $(3,2,1)$ auftritt. Nach den Pfadregeln folgt also $\mathbb{P}(X = 4) = 6 \cdot \left(\frac{1}{3}\right)^3 = \frac{2}{9}$. Hieran erkennt man, dass allgemein die Reihenfolge, in der die Teilchen in die einzelnen Fächer fallen, beachtet werden muss. Wegen $\mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = 1$ folgt nun $\mathbb{P}(X = 3) = \frac{4}{9}$ und damit $\mathbb{E}(X) = \frac{26}{9}$.

Wohingegen man für die Fälle $n = 3$ und $n = 4$ noch mit reduzierten Baumdiagrammen arbeiten kann, ist ein solches Vorgehen für den Fall $n \geq 5$ kaum praktikabel, und jetzt zeigt sich die Kraft der Kombinatorik. Um etwa im Fall $n = 5$ die Wahrscheinlichkeit $\mathbb{P}(X = 4)$ zu berechnen, kann folgendermaßen argumentiert werden: Zuerst sind drei der fünf Fächer auszuwählen, die dann mit jeweils einem Teilchen besetzt werden. Dafür gibt es $\binom{5}{3}$ Möglichkeiten. Da bei dieser Modellierung die Reihenfolge der Besetzungen notiert wird (siehe oben im Fall $n = 3$), tritt der Faktor $3!$ auf. Für das vierte Teilchen gibt es noch drei Möglichkeiten. Nach dem fundamentalen Zählprinzip folgt somit

$$\mathbb{P}(X = 4) = \binom{5}{3} \cdot 3! \cdot 3 \cdot \frac{1}{5^4}.$$

Der Schritt zur Verallgemeinerung ist dann nicht mehr weit. Insgesamt ergeben sich viele Vernetzungen und Anlässe zum Argumentieren.

Mit GeoGebra lassen sich Stabdiagramme der Verteilung von X dynamisch darstellen (s. Abb. 2). Zusätzlich können Kenngrößen wie der Erwartungswert und die Standardabweichung von X ausgegeben werden. Sogar eine Kontrolle der Normierungsgsbedingung $\sum_{k=2}^{n+1} \mathbb{P}(X = k) = 1$ kann mithilfe von GeoGebra-CAS erfolgen.

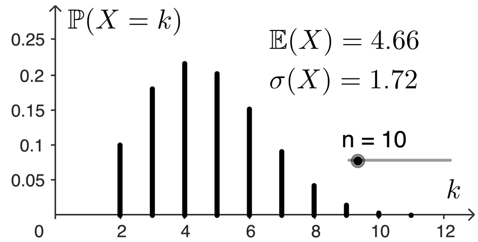

Abb. 2: Stabdiagramm der Verteilung von X für $n = 10$

Was passiert, wenn die Gleichverteilungsannahme fallen gelassen wird? Wie verändert sich dann der Graph in Abb. 2? Mit diesen anspruchsvollen Fragen könnte der Übergang eingeleitet werden. Es ist auch möglich, einen Wettbewerb zu initiieren, etwa durch den folgenden Impuls: Wir haben $\mathbb{P}(X = 3) = \frac{1}{3}$ für den Fall $n = 3$ erhalten. Versuchen Sie, diesen Wert zu unterschreiten, wenn die Gleichverteilungsannahme fallengelassen wird.

Natürlich muss jetzt präzisiert werden, was es heißt, „die Gleichverteilungsannahme fallenzulassen“. Wir nummerieren die n Fächer gedanklich von 1 bis n und bezeichnen mit p_j die Wahrscheinlichkeit, dass ein Teilchen in das Fach j fällt, $j = 1, \ldots, n$. Dabei setzen wir $p_j > 0$ für jedes j voraus, denn sonst könnte man sich von vornherein auf weniger als n Fächer beschränken. Die Gleichverteilungsannahme bedeutet jetzt speziell $p_j = \frac{1}{n}$ für jedes $j = 1, \ldots, n$.

Es ist wichtig, dass Schülerinnen und Schüler vor der Theorie eigene Erfahrungen sammeln und Vermutungen aufstellen, die dann überprüft werden. An dieser Stelle können auch gewinnbringend Simulationen eingesetzt werden. Abb. 3 zeigt eine einfache Umsetzung mit GeoGebra für den Fall $n = 10$.

Abb. 3: einfache Simulation mit GeoGebra

Mit dem Befehl *ZufallszahlDiskret(Liste1,Liste2)* wird die zufällige Fächerverteilung realisiert. Liste1 symbolisiert die 10 Fächer, Liste2 enthält die zugehörigen Wahrscheinlichkeiten. Mit der Liste l2 wird eine Gleichverteilung simuliert, bei Liste l3 gilt $p_1 = p_2 = \frac{3}{14}$ und $p_i = \frac{1}{14}$ für $i = 3, \ldots, 10$. Im zweiten Fall erhält also jedes der Fächer 1 und 2 eine dreimal so große Wahrscheinlichkeit wie jedes andere Fach. In den Listen l2 und l3 ist jeweils die erste Stelle zu suchen, an der eine Zahl zum ersten Mal doppelt vorkommt. Damit kann jeweils die zugehörige Häufigkeitsverteilung ermittelt werden. Dieses behutsame Vorgehen ist für das Verständnis wichtig, die Eigenaktivität wird gefördert. Durch Zusammenfassen der Ergebnisse kann eine große Anzahl an Daten generiert werden, um damit Schätzwerte für die Wahrscheinlichkeiten $\mathbb{P}(X = k)$ zu erhalten.

Es sollte entdeckt werden: Frühe Kollisionen werden im Fall der Gleichverteilungsannahme wahrscheinlicher und späte unwahrscheinlicher. Ist das immer so? Falls ja, kann das begründet werden? Wir müssen Schülerinnen und Schüler immer wieder Gelegenheiten geben, Fragen zu stellen.

Im Folgenden wird die Theorie vorgestellt. Im Hinblick auf eine Umsetzung im Unterricht betrachten wir nur die Fälle $n = 2$ und $n = 3$. Anhand des letzten Falls kann man aber selbst auf Schulniveau einsehen, *warum* die Wahrscheinlichkeit für mindestens einen Mehrfachgeburtstag unter k Personen minimal wird, wenn die Geburtstage gleich wahrscheinlich sind, siehe Abschn. 4.

2 Der Fall $n = 2$

Wir setzen kurz $p := p_1$, sodass $p_2 = 1 - p$ gilt. Da im Fall $n = 2$ die erste Kollision spätestens mit dem dritten Teilchen erfolgt, nimmt die Zufallsgröße X die möglichen Werte 2 und 3 an. Für diesen Fall ist folgende Aufgabe denkbar:

Aufgabe 1: Berechnungen, p beliebig
Zeigen Sie:

a) $\mathbb{P}(X = 2) = p^2 + (1-p)^2$,

b) $\mathbb{P}(X = 3) = 2p(1-p)$,

c) $\mathbb{E}(X) = 2 + 2p(1-p)$.

Zu a) Es müssen die beiden Fälle unterschieden werden, dass das erste Teilchen in Fach 1 oder Fach 2 fällt. Damit bereits beim zweiten Teilchen eine Kollision erfolgt, muss dieses in das gleiche Fach wie das erste Teilchen fallen.

Zu b) Man kann zum einen Teil a) und die Gleichung $\mathbb{P}(X = 3) = 1 - \mathbb{P}(X = 2)$ verwenden. Man kann aber auch direkt argumentieren: Die erste Kollision erfolgt genau dann mit dem dritten Teilchen, wenn die beiden ersten Teilchen in verschiedene Fächer fallen. Dafür gibt es die beiden Fälle $(1,2)$ und $(2,1)$. Die Wahrscheinlichkeit für $(1,2)$ ist $p(1-p)$, und die für $(2,1)$ ist $(1-p)p$.

Zu c) Die Darstellung für $\mathbb{E}(X)$ folgt direkt aus $\mathbb{E}(X) = 2 \cdot \mathbb{P}(X = 2) + 3 \cdot \mathbb{P}(X = 3)$ sowie a) und b).

Da die Funktion $p \mapsto p(1-p)$ ihr Maximum $\frac{1}{4}$ für $p = \frac{1}{2}$ annimmt, wartet man unter der Gleichverteilung im Mittel am längsten auf die erste Kollision. Zudem wird klar, dass kleine Abweichungen von p zum Laplace-Fall $p = \frac{1}{2}$ den maximalen Erwartungswert 2.5 unter der Gleichverteilungsannahme kaum verringern. Gilt $0.4 \leq p \leq 0.6$, so ist $\mathbb{E}(X) \geq 2.48$. Aussage b) zeigt, dass die Wahrscheinlichkeit $\mathbb{P}(X = 3)$ für eine spätestmögliche Kollision ebenfalls unter der Gleichverteilungannahme *maximal* wird, und die wegen $\mathbb{P}(X = 2) + \mathbb{P}(X = 3) = 1$ dazu komplementäre Aussage a) bedeutet, dass die Wahrscheinlichkeit für eine frühestmögliche Kollision unter dieser Annahme *minimal* wird.

3 Der Fall $n = 3$

In diesem Fall nimmt X die möglichen Werte 2, 3 und 4 an, und es ist ein ganzes Arsenal an Aufgaben möglich. Wir starten mit dem Spezialfall, dass zwei der drei Fächer als gleichwahrscheinlich angenommen werden und treffen für jedes p mit $0 < p < \frac{1}{2}$ die Festsetzung $p_1 = p_2 := p$. Damit ergibt sich p_3 zu $1 - 2p$. Die Wahrscheinlichkeiten $\mathbb{P}(X = k)$ sind dann Funktionen von p, und wir schreiben

$$f_k(p) := \mathbb{P}(X = k), \quad k = 2, 3, 4.$$

Aufgabe 2: Spezialfall für $k = 2$; Minimum gesucht
Zeigen Sie: Es gilt

$$f_2(p) = 6p^2 - 4p + 1, \tag{1}$$

und die Funktion f_2 nimmt ihr Minimum an der Stelle $p = \frac{1}{3}$ an.

Das Ereignis $\{X = 2\}$ tritt ein, wenn die beiden ersten Teilchen in das gleiche Fach fallen. Unterscheidet man die Fälle $(1,1)$, $(2,2)$ und $(3,3)$, so folgt $f_2(p) = p^2 + p^2 + (1-2p)^2$, was mit (1) übereinstimmt. Die zweite Behauptung folgt durch Nullsetzen der Ableitung. Die Wahrscheinlichkeit für eine Kollision nach nur zwei Teilchen wird also im Fall der Gleichverteilung minimal.

Aufgabe 3: Spezialfall für $k = 4$; Maximum gesucht
Zeigen Sie: Es gilt

$$f_4(p) = 6p^2 - 12p^3, \tag{2}$$

und die Funktion f_4 nimmt ihr Maximum für $p = \frac{1}{3}$ an.

Hier sollten die Schülerinnen und Schüler erkennen, dass es für das Eintreten des Ereignisses $\{X = 4\}$ die 6 verschiedenen Fälle $(1,2,3)$, $(1,3,2)$, $(2,1,3)$, $(2,3,1)$, $(3,1,2)$ und $(3,2,1)$, also alle sechs Permutationen der Zahlen 1, 2 und 3, gibt. Da die Multiplikation kommutativ ist, besitzt jeder dieser Fälle nach der ersten Pfadregel die Wahrscheinlichkeit $p^2(1-2p)$, was zu $f_4(p) = 6p^2(1-2p)$ und damit zu (2) äquivalent ist. Die zweite Behauptung folgt dann mit einer Kurvendiskussion.

Wegen $\mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = 1$ gilt $f_3(p) = 1 - f_2(p) - f_4(p)$, und eine direkte Rechnung ergibt $f_3(p) = 12p^3 - 12p^2 + 4p$.

Es gilt $f_3'(p) = 36\left(p - \frac{1}{3}\right)^2$. Somit ist f_3 streng monoton wachsend und besitzt wegen $f_3''(\frac{1}{3}) = 0$ einen Sattelpunkt an der Stelle $p = \frac{1}{3}$. Abb. 4 zeigt Schaubilder der Funktionen f_2, f_3 und f_4. Die gepunktete vertikale Linie geht durch den Punkt $(\frac{1}{3}, 0)$.

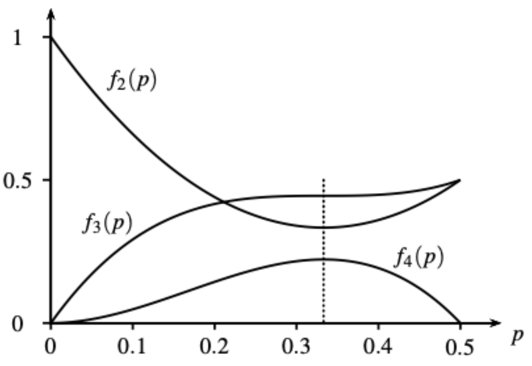

Abb. 4: Graphen der Funktionen f_2, f_3 und f_4

Aufgabe 4: Erwartungswert für den Spezialfall
Zeigen Sie, dass der Erwartungswert von X durch

$$e(p) = -12p^3 + 4p + 2$$

gegeben ist. Für welchen Wert von p wird $e(p)$ maximal?

Nach der Berechnungsformel für den Erwartungswert gilt $e(p) = \mathbb{E}(X) = 2f_2(p) + 3f_3(p) + 4f_4(p)$. Einsetzen und Zusammenfassen von Termen mit gleicher Potenz von p liefert die Behauptung. Eine Kurvendiskussion zeigt, dass die Funktion e an der Stelle $\frac{1}{3}$ einen Hochpunkt besitzt. Somit wird der Erwartungswert der Anzahl der Teilchen bis zur ersten Kollision im Fall der Gleichverteilung maximal.

Wir wenden uns nun dem allgemeinen Fall von drei Fächern mit den Wahrscheinlichkeiten p_1, p_2 und p_3 zu.

Aufgabe 5: $\mathbb{P}(X = 2)$ und Minimum – allgemein
Zeigen Sie, dass

$$\mathbb{P}(X = 2) = p_1^2 + p_2^2 + p_3^2$$

gilt und $\mathbb{P}(X = 2)$ *nur* unter dem Laplace-Modell $p_1 = p_2 = p_3 = \frac{1}{3}$ minimal wird.

Der erste Teil der Aufgabe erfordert die gleiche Einsicht, wie sie nach Aufgabe 2 beschrieben ist. Es gilt $\mathbb{P}(X = 2) = p_1^2 + p_2^2 + p_3^2$. Interessant ist natürlich der Aspekt, $\mathbb{P}(X = 2)$ als Funktion von p_1, p_2 und p_3 anzusehen und zu minimieren. Wegen $p_3 = 1 - p_1 - p_2$ ist $\mathbb{P}(X = 2)$ eine Funktion von p_1 und p_2. Uns ist bewusst, dass Funktionen von zwei Veränderlichen in keinem Lehrplan auftreten. Gleichwohl lassen sich mit dem sinnvollen Einsatz digitaler Medien an dieser Stelle Methoden anwenden, die in den Lehrplänen an exponierter Stelle stehen.

Der Graph einer Funktion f mit $f(x,y,z) = x^2 + y^2 + z^2$ unter der Nebenbedingung $x + y + z = 1$ kann zuerst grafisch veranschaulicht werden. Dies geschieht z.B. beim Einsatz von GeoGebra durch den Befehl $f(x,y,1-x-y)$, wenn vorher $f(x,y,z) := x^2 + y^2 + z^2$ definiert wurde. Schon ist zu vermuten, dass genau ein Tiefpunkt $P(\frac{1}{3}|\frac{1}{3}|\frac{1}{3})$ existiert. Gewissheit wird erlangt, wenn der Graph sowie die Funktionsgleichung von g mit $g(x,y) := f(x + \frac{1}{3}, y + \frac{1}{3})$ untersucht werden (s. Abb. 5).

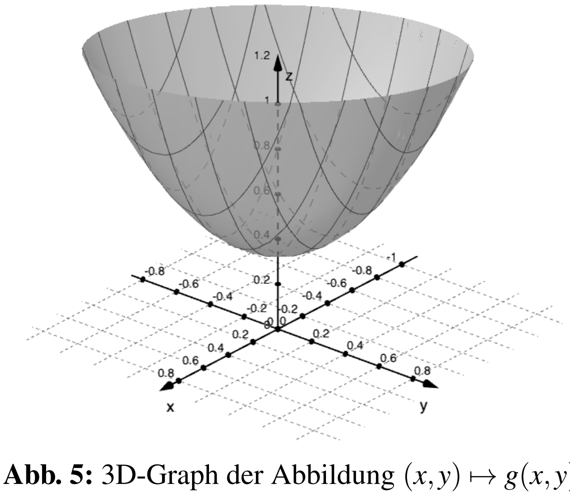

Abb. 5: 3D-Graph der Abbildung $(x,y) \mapsto g(x,y)$

GeoGebra liefert sofort $g(x,y) = x^2 + y^2 + 2xy + \frac{1}{3}$. Durch die Transformation von f zu g hat sich eine zur z-Achse achsensymmetrische Funktion ergeben ($g(x,y) = g(-x,-y)$), die (nur) an der Stelle $(0,0)$ ein globales Minimum besitzt. Für unser Ausgangsproblem bedeutet dies, dass die Wahrscheinlichkeit $\mathbb{P}(X = 2)$ (nur) im Laplace-Fall minimal wird.

Durch den Einsatz digitaler Medien rücken sowohl grafische Möglichkeiten als auch der Objektcharakter von Funktionen immer stärker in den Fokus. Es gibt in Schulbüchern und auch in Prüfungsaufgaben zunehmend Aufgaben, die Verschiebungen von Graphen verlangen – leider oft ohne tieferen Sinn. In unserem Fall wird der Sinn deutlich: Bei der vorgenommenen Transformation verändert sich der Wert des Minimums nicht. Es entsteht ein Term, der eine Abschätzung erleichtert, da die Terme x^2, y^2, $2xy$ für alle $x,y > 0$ positiv sind und sich nur für $x = y = 0$ der Wert null als Summe ergibt.

Aufgabe 6: $\mathbb{P}(X = 4)$ und Maximum – allgemein
Zeigen Sie:

$$\mathbb{P}(X = 4) = 6p_1p_2p_3.$$

Hinweis: Das Ereignis $\{X = 4\}$ besagt, dass die ersten drei Teilchen in drei verschiedene Fächer fallen.

Hier argumentiert man wie im Absatz nach Formel (2). Jeder der sechs Fälle (i,j,k) mit drei verschiedenen Werten 1, 2 oder 3 für i, j und k hat nach der ersten Pfadregel die Wahrscheinlichkeit $p_1p_2p_3$.

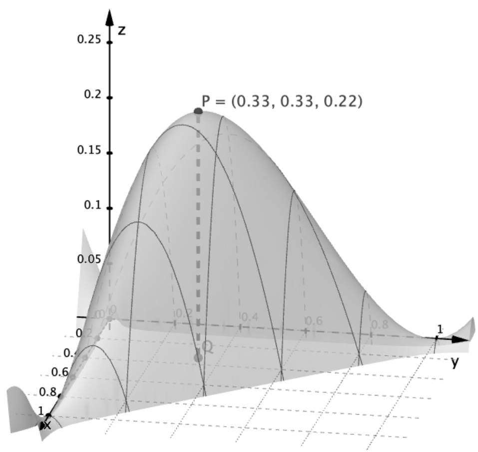

Abb. 6: Graph der Funktion $(x,y) \mapsto 6xy(1-x-y)$

Aufgrund dieses Ergebnisses stellt sich die Frage, ob die Funktion $(p_1,p_2,p_3) \mapsto p_1p_2p_3$ unter der Nebenbedingung $p_1 + p_2 + p_3 = 1$ im Gleichverteilungsfall $p_1 = p_2 = p_3$ maximal wird. Das Produkt $p_1p_2p_3$ beschreibt der Rauminhalt eines Quaders mit den Seitenlängen p_1, p_2 und p_3. Wie groß kann dieser Rauminhalt maximal werden, wenn die Summe der Seitenlängen gleich eins ist? Wie die folgenden Ausführungen zeigen, ist die Antwort auf diese Frage: Für den Würfel mit Seitenlänge $\frac{1}{3}$.

Der Graph von f mit $f(x,y) := 6 \cdot xy(1-x-y)$ liefert die Vermutung, dass $P(\frac{1}{3}|\frac{1}{3}|\frac{2}{9})$ für $0 \leq x,y \leq 1$ der Hochpunkt ist (s. Abb. 6).

Der Ansatz einer geschickten Verschiebung liefert wie nach Aufgabe 5 eine Begründung. So produziert GeoGebra mit der Definition $g(x,y) := f(x+\frac{1}{3}, y+\frac{1}{3})$ für g den Funktionsterm

$$\frac{2}{9} - (6x^2y + 2x^2 + 6xy^2 + 2xy + 2y^2),$$

der für alle $0 \leq x,y \leq 1$ nur für $x = y = 0$ den maximalen Wert $\frac{2}{9}$ annimmt. Aus diesem Grund ist die Wahrscheinlichkeit $\mathbb{P}(X = 4)$ für eine spätestmögliche erste Kollision nur im Fall der Laplace-Verteilung maximal.

Wieder zeigt eine (nicht vom Himmel fallende!) geschickte Verschiebung eine Termstruktur, die den maximalen Wert sichtbar werden lässt. Insgesamt liegt ein schönes Zusammenspiel von Graph, einer Transformation und einer Termstrukturerkennung vor.

Aufgabe 7: $\mathbb{E}(X)$ und Maximum – allgemein
Zeigen Sie:

$$\mathbb{E}(X) = 3 - p_1^2 - p_2^2 - p_3^2 + 6p_1p_2p_3. \qquad (3)$$

Begründen Sie, dass $\mathbb{E}(X)$ genau dann maximal wird, wenn die Gleichverteilung $p_1 = p_2 = p_3 = \frac{1}{3}$ vorliegt. Wie groß ist dieser maximale Wert?

Mit Aufgabe 5 und Aufgabe 6 sowie der Berechnungsformel für den Erwartungswert folgt zunächst

$$\begin{aligned}\mathbb{E}(X) &= 2 \cdot \mathbb{P}(X = 2) + 3 \cdot \mathbb{P}(X = 3) + 4 \cdot \mathbb{P}(X = 4)\\ &= 2(p_1^2 + p_2^2 + p_3^2) + 3 \cdot \mathbb{P}(X = 3) + 24p_1p_2p_3.\end{aligned}$$

Wegen $\mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 4) = 1$ gilt

$$\mathbb{P}(X = 3) = 1 - (p_1^2 + p_2^2 + p_3^2) - 6p_1p_2p_3,$$

sodass sich das Ergebnis durch Einsetzen ergibt.

Wir schätzen den Erwartungswert in (3) nach oben ab, indem wir die Quadratsumme $p_1^2 + p_2^2 + p_3^2$ nach unten und den Term $6p_1p_2p_3$ nach oben abschätzen. Nach Aufgabe 5 nimmt die Quadratsumme nur im Laplace-Fall den kleinstmöglichen Wert $\frac{1}{3}$ an. Des Weiteren haben wir gesehen, dass das Produkt $p_1p_2p_3$ nur im Laplace-Fall den maximalen Wert $\frac{1}{27}$ annimmt. Es folgt

$$\mathbb{E}(X) \leq 3 - \frac{1}{3} + \frac{6}{27} = \frac{26}{9} \approx 2.889,$$

wobei die obere Schranke nur im Laplace-Fall angenommen wird.

Anmerkung: Hintergrund der Abschätzungen in den Aufgaben 6 und 7 ist im Wesentlichen die allgemein für n positive Zahlen $x_1,\ldots,x_n$ geltende Ungleichung

$$\left(x_1\cdot\ldots\cdot x_n\right)^{1/n}\leq\bar{x}:=\frac{1}{n}\left(x_1+\ldots+x_n\right)\qquad(4)$$

zwischen dem geometrischen und dem arithmetischen Mittel. Dabei tritt das Gleichheitszeichen nur dann ein, wenn $x_1=\ldots=x_n$ gilt. Der Beweis von (4) kann allein mithilfe der für jedes t geltenden Ungleichung $t+1\leq e^t$ erfolgen, bei der Gleichheit nur für $t=0$ gilt. Mithilfe dieser Ungleichung folgt

$$\frac{x_j}{\bar{x}}\leq\exp\left(\frac{x_j}{\bar{x}}-1\right),\quad j\in\{1,\ldots,n\}.$$

Wegen $e^{a+b}=e^a e^b$ und $(x_1+\ldots+x_n)/\bar{x}=n$ liefert Multiplikation dieser Ungleichungen

$$\frac{1}{\bar{x}^n}\left(x_1\cdot\ldots\cdot x_n\right)\leq e^0=1,$$

woraus (4) sowie der Zusatz über das Eintreten des Gleichheitszeichens folgen. Wegen $p_1+p_2+p_3=1$ gilt also insbesondere $(p_1p_2p_3)^{1/3}\leq\frac{1}{3}$.

4 Fächerwahrscheinlichkeiten mitteln vergrößert $\mathbb{P}(X>k)$

In der Einleitung wurde hervorgehoben, dass bei Abweichung von der Gleichverteilung die Wahrscheinlichkeit, dass von k Personen mindestens zwei am gleichen Tag Geburtstag haben, für jedes $k\in\{2,\ldots,365\}$ größer wird. Das hierzu komplementäre Ereignis stellt sich in der Form $\{X>k\}$ dar und besagt (mit allgemeinem n anstelle von 365 und neutraler Formulierung), dass die ersten k Teilchen in lauter verschiedene der n Fächer fallen.

In diesem Abschnitt geht es zum einen um die Berechnung von $\mathbb{P}(X>k)$ und zum anderen darum, diese Wahrscheinlichkeit durch eine geeignete Modifikation der Fächerwahrscheinlichkeiten $p_1,\ldots,p_n$ zu vergrößern, solange keine Gleichverteilung vorliegt und damit mindestens ein p_j verschieden von $\frac{1}{n}$ ist. Indem man derartige Modifikationen mehrfach vornimmt, nähert man sich der Gleichverteilung immer mehr an. Ein Stetigkeits- und ein Kompaktheitsargument zeigen dann, dass $\mathbb{P}(X>k)$ für jedes $k\in\{2,\ldots,n\}$ nur unter der Gleichverteilung maximal werden. Da man den Erwartungswert von X als Summe obiger Wahrscheinlichkeiten schreiben kann, wird auch $\mathbb{E}(X)$ nur unter der Gleichverteilung am größten.

Wir möchten diese Überlegungen zunächst anhand des Spezialfalls $n=3$ konkretisieren und schreiben hierzu den Erwartungswert von X in der Form

$$\begin{aligned}\mathbb{E}(X)&=2\mathbb{P}(X=2)+3\mathbb{P}(X=3)+4\mathbb{P}(X=4)\\&=2\big(\mathbb{P}(X=2)+\mathbb{P}(X=3)+\mathbb{P}(X=4)\big)\\&\quad+\big(\mathbb{P}(X=3)+\mathbb{P}(X=4)\big)+\mathbb{P}(X=4)\\&=2+\mathbb{P}(X>2)+\mathbb{P}(X>3).\end{aligned}\qquad(5)$$

Für die hier auftretenden Wahrscheinlichkeiten gelten

$$\mathbb{P}(X>2)=2\big(p_1p_2+p_1p_3+p_2p_3\big),\qquad(6)$$
$$\mathbb{P}(X>3)=6p_1p_2p_3.\qquad(7)$$

Wegen $\mathbb{P}(X>3)=\mathbb{P}(X=4)$ folgt dabei die zweite Aussage aus Aufgabe 6. Um (6) zu zeigen, beachte man, dass das Ereignis $\{X>2\}$ bedeutet, dass die beiden ersten Teilchen in verschiedene Fächer fallen, wofür es die Fälle $(1,2)$, $(2,1)$, $(1,3)$, $(3,1)$, $(2,3)$ und $(3,2)$ gibt. Addition der zugehörigen Wahrscheinlichkeiten liefert (6).

Da die Wahrscheinlichkeiten $\mathbb{P}(X>2)$ und $\mathbb{P}(X>3)$ und somit auch der Erwartungswert von X nicht davon abhängen, in welcher Weise wir die drei Fächer nummerieren, sei $p_1\leq p_2\leq p_3$ vereinbart. Dann liegt genau dann keine Gleichverteilung vor, wenn $p_1<p_3$ gilt. Jetzt präzisieren wir, was wir mit einem wohl auf Munford (1977) zurückgehenden Kunstgriff, die Fächerwahrscheinlichkeiten geeignet zu *modifizieren*, meinen. Wir betrachten nämlich neue Wahrscheinlichkeiten q_1,q_2 und q_3 für die Fächer, und zwar setzen wir

$$q_1:=\frac{p_1+p_3}{2},\;\;q_2:=p_2,\;\;q_3:=\frac{p_1+p_3}{2}.$$

Im Spezialfall $p_1=0.2$, $p_2=0.3$ und $p_3=0.5$ würde sich also $q_1=0.35$, $q_2=0.3$ und $q_3=0.35$ ergeben. Wir werden jetzt sehen, dass die Wahrscheinlichkeit $\mathbb{P}(X>2)$ größer wird, wenn sich die Wahrscheinlichkeiten der Fächer von p_1,p_2,p_3 zu q_1,q_2,q_3 ändern. Wahrscheinlichkeiten, die sich auf die Fächerwahrscheinlichkeiten q_1,q_2,q_3 beziehen, notieren wir in der Form $\mathbb{P}_*$ anstelle von $\mathbb{P}$.

Ersetzt man in (6) und (7) p_j durch q_j $(j=1,2,3)$, so liefert eine kurze Rechnung

$$\begin{aligned}\mathbb{P}_*(X>2)&=\mathbb{P}(X>2)+\frac{1}{2}\big(p_3-p_1\big)^2,\\\mathbb{P}_*(X>3)&=\mathbb{P}(X>3)+\frac{3}{2}p_2\big(p_3-p_1\big)^2.\end{aligned}$$

Damit gelten $\mathbb{P}_*(X > 2) > \mathbb{P}(X > 2)$ sowie $\mathbb{P}_*(X > 3) > \mathbb{P}(X > 3)$. Wegen (5) vergrößert sich auch $\mathbb{E}(X)$, wenn wir von p_1, p_2, p_3 zu q_1, q_2, q_3 übergehen.

Solange keine Gleichverteilung vorliegt, können wir also $\mathbb{E}(X)$ vergrößern, indem wir durch Mittelung der beiden größten und kleinsten Fach-Wahrscheinlichkeiten „näher an die Gleichverteilung heranrücken“. Im obigen Zahlenbeispiel $p_1 = 0.2$, $p_2 = 0.3$ und $p_3 = 0.5$ haben wir nach der ersten solchen Mittelung die sortierten Wahrscheinlichkeiten 0.3, 0.35 und 0.35 erhalten. Nach einer weiteren Mittelung gelangen wir zu den der Größe nach sortierten Wahrscheinlichkeiten 0.325, 0.325 und 0.35 und rücken wieder etwas näher an die Gleichverteilung $p_1 = p_2 = p_3 = \frac{1}{3}$ heran. Diese Gleichverteilung würden wir im Grenzwert $\ell \to \infty$ erhalten, wenn wir ℓ Iterationen des oben definierten „Mittelungs-Algorithmus“ vornehmen.

Im allgemeinen Fall von n Fächern gilt

$$\mathbb{P}(X > k) = k! \sum_{1 \le i_1 < \ldots < i_k \le n} p_{i_1} p_{i_2} \cdot \ldots \cdot p_{i_k}$$

(s. z.B. Henze (2015), Abschn. 3). Es müssen ja die Nummern $i_1, \ldots, i_k$ der Fächer ausgewählt werden, in die die k Teilchen fallen sollen. Bei gegebener Auswahl dieser Nummern gibt es $k!$ Reihenfolgen, diese Fächer zu besetzen. Wir nehmen an, dass $p_1 \le \ldots \le p_n$ gilt, denn obige Wahrscheinlichkeit hängt nicht von der Nummerierung der Fächer ab. Außerdem gelte $p_1 < p_n$, was gleichbedeutend damit ist, dass keine Gleichverteilung vorliegt. In Henze (2015) wurde gezeigt, dass sich die Wahrscheinlichkeit $\mathbb{P}(X > k)$ vergrößert, wenn man von den Fächerwahrscheinlichkeiten $p_1, \ldots, p_n$ zu Fächerwahrscheinlichkeiten $q_1, \ldots, q_n$ übergeht, die durch

$$q_1 := q_n := \frac{p_1 + p_n}{2}, \quad q_j := p_j \text{ für } j = 2, \ldots, n-1,$$

definiert sind. Da $\mathbb{P}(X > 2)$ eine stetige Funktion von $p_1, \ldots, p_n$ und der Definitionsbereich dieser Funktion die kompakte Menge $\{(p_1, \ldots, p_n) : 0 \le p_j \le 1 \text{ für } j = 1, \ldots, n; p_1 + \ldots + p_n = 1\}$ ist, wird die Wahrscheinlichkeit $\mathbb{P}(X > 2)$ (nur) im Fall der Gleichverteilung maximal. In Verallgemeinerung von (5) gilt

$$\mathbb{E}(X) = 2 + \sum_{k=2}^{n} \mathbb{P}(X > k),$$

und somit wird auch der Erwartungswert von X (nur) im Laplace-Fall maximal. Dieser Wert ist robust gegenüber Abweichungen von der Gleichverteilungsannahme. So liefert z. B. eine Simulation mit den deutlich von einer Gleichverteilung abweichenden Wahrscheinlichkeiten $p_j = 2j/(365 \cdot 366)$, $j \in \{1, \ldots, 365\}$, und 10000 Wiederholungen für $n = 365$ den Schätzwert 21.5. Im Gleichverteilungsfall gilt (auf eine Nachkommastelle gerundet) $\mathbb{E}(X) = 24.6$, s. auch Barth und Haller (2013), S. 26.

5 Geburtstagsproblem und Minimum

Die Überlegungen des vorigen Abschnitts resultieren in der Erkenntnis, dass für jedes k mit $k \in \{2, \ldots, n\}$ die Wahrscheinlichkeit $\mathbb{P}(X > k)$ für den Fall der Gleichverteilung maximal und damit die Wahrscheinlichkeit des komplementären Ereignisses $\{X \le k\}$ für diesen Fall minimal wird. Abb. 7 illustriert diesen Sachverhalt anhand des Geburtstagsproblems, d.h. des Falls $n = 365$.

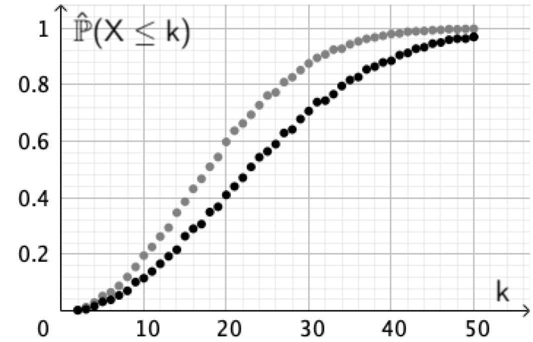

Abb. 7: $\mathbb{P}(X \le k)$ geschätzt ($n = 365$, 5000 Simulationen), mit und ohne Gleichverteilung

Zur Anfertigung von Abb. 7 wurde mit GeoGebra für jedes k mit $2 \le k \le 50$ die Situation des Geburtstagsproblems mit k Personen simuliert. Zu jedem Abszissenwert k gibt der Ordinatenwert des gezeichneten Punktes die relative Häufigkeit aller Fälle an, in denen mindestens ein Doppelgeburtstag auftrat. Diese relative Häufigkeit ist somit ein mit $\hat{\mathbb{P}}(X \le k)$ bezeichneter Schätzwert für die Wahrscheinlichkeit $\mathbb{P}(X \le k)$. Dabei wurde für die schwarzen Punkte eine Gleichverteilung zugrunde gelegt. Bei den in Abb. 7 grau markierten Punkten haben alle Tage ab dem 183. Tag eine 10-mal so große Wahrscheinlichkeit gezogen zu werden wie die ersten 182 Tage. Deutlich zu erkennen ist das bekannte Resultat, nach dem ab einer Zahl von 23 Personen darauf gewettet werden kann, dass mindestens zwei von ihnen am gleichen Tag Geburtstag haben. Für die modifizierte Verteilung sind es nur 18 Personen. Im Unterricht sollte zuerst mit Realdaten gearbeitet werden, z. B. mit den Geburtstagen der Schülerinnen und Schüler. Auch die einfach zu erhaltenen Geburtstage von Fußballspielerinnen und -spielern eignen sich sehr gut. Die Umsetzung mit GeoGebra gelingt erstaunlich einfach (s. Abb. 8). Der Befehl *DotPlot* ermöglicht eine schöne Visualisierung der Fächerbesetzungen (s. Abb. 9). Mehr noch: Hiermit lassen sich viele Erkundungen durchführen und erste Schätzungen für die gesuchten Wahrscheinlichkeiten ermitteln.

Sämtliche GeoGebra-Dateien können unter der Adresse https://www.geogebra.org/m/vjxrkurm bearbeitet bzw. heruntergeladen werden.

Abb. 8: einfache Umsetzung mit GeoGebra

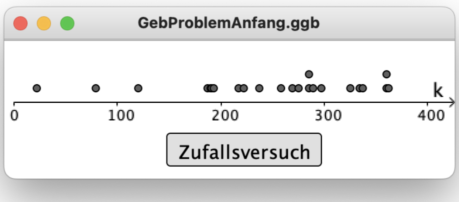

Abb. 9: Punkteplot einer zugehörigen Simulation

6 Fazit

Im Stile einer typischen Schulbuchaufgabe fing alles mit zwei Fächern an, die als gleich wahrscheinlich angenommen wurden. In einem allgemeinbildenden Mathematikunterricht sollten Annahmen aber auch hinterfragt werden, mit lohnenswertem Ergebnis! Es zeigt sich nämlich, dass die Gleichverteilung eine Extremaleigenschaft besitzt und damit etwas ganz Besonderes ist. Ein Abweichen davon kann schnell kompliziert, aber auch spannend werden, denn dann treffen sich Stochastik und Analysis. Andererseits zeigt sich auch, dass die Annahme einer Gleichverteilung beim Geburtstagsproblem gar nicht fragwürdig ist, denn interessierende Wahrscheinlichkeiten sind robust gegenüber realistischen Abweichungen von einer solchen Annahme.

Danksagung: Wir danken den Gutachtern sowie Joachim Engel für wertvolle Hinweise.

Literatur

Arnold, M., Glaß, W. (2013): Simple Approximation Formulas for the Birthday Problem. *American Mathematical Monthly, 120*(7), 645–648.

Barth, F., Haller, R. (2012): Besetzungen und Geburtstage. *Stochastik in der Schule, 32*(3), 20–27.

Barth, F., Haller, R. (2013): Gemeinsame Geburtstage. *Stochastik in der Schule, 33*(1), 25–32.

Gorenflo, S., Sauer, M. (2018): Ein Problem beim Geburtstagsproblem. *Stochastik in der Schule, 38*(1), 21–25.

Henze, N. (1995): Erstmals im Lotto dieselbe Zahlenreihe – eine Sensation? *Der Mathematisch-Naturwissenschaftliche Unterricht (MNU), 48*(8), 456–457.

Henze, N. (2015): Stochastische Extremwertprobleme im Fächer-Modell I: Minima von Wartezeiten und Kollisionsprobleme. *Stochastik in der Schule, 35*(3), 24–30.

Knuth, D. E. (1998): The Art of Computer Programming. Volume 3. Sorting and Searching, 2. Auflage. Reading, Massachusetts: Addison-Wesley.

Munford, A.G. (1977): A Note on the Uniformity Assumption in the Birthday Problem. *The American Statistician, 31*(3), 119.

Riehl, G. (2006): Neues zum Geburtstagsproblem. *Der Mathematisch-Naturwisseschaftliche Unterricht (MNU), 59*(7), 439–444.

Riehl, G. (2014): Alte Geburtstagsprobleme – neu gelöst. *Mathematische Semesterberichte, 61*(2), 215–232.

Schrage, G. (1990): Ein Geburtstagsproblem. *Mathematische Semesterberichte, 37*(2), 251–257.

Von Mises, R. (1939): Über Aufteilungs- und Besetzungswahrscheinlichkeiten. *Revue de la Faculté des Sciences de l'Université d'Istanbul, 4*, 145–163.

Anschrift der Verfasser:

Prof. i.R. Dr. Norbert Henze
KIT Distinguished Senior Fellow
Institut für Stochastik
Karlsruher Institut für Technologie (KIT)
Englerstr. 2
76131 Karlsruhe
`Henze@kit.edu`

Reimund Vehling
`vehling@icloud.com`