Manfred Borovcnik und Martin Schenk, Klagenfurt

Simulationen im Stochastik-Unterricht

Der Stellenwert von Simulationen zur Stützung von Einsicht und für den Aufbau von Verständnis schwieriger Konzepte in der Stochastik ist unbestritten. Darüber hinaus kann man den Gedanken von Modellbildung in den Unterricht ernsthafter einbinden, wenn man die komplexen Berechnungen durch Simulation ersetzt und die Unterweisung darauf fokussiert, welchen Voraussetzungen das eingesetzte Modell genügen muss und wie oder wie sehr – und warum – man in realen Situationen davon abweichen kann. Neben Simulationen sind animierte didaktische Sequenzen manchmal sehr hilfreich, um den Stellenwert von Einflussparametern einzuschätzen oder Invarianzen und damit wesentliche Eigenheiten von Begriffen zu erkennen. Eine Tabellenkalkulation wie EXCEL bietet hier eine Fülle von Möglichkeiten, deren besonderer Vorteil auch in der dadurch ermöglichten Interaktivität liegt. Das wird natürlich in einem Vortrag besser erlebbar als in einem Aufsatz.

1. Vorbemerkungen

Die Stochastik gilt als schwer. Betrachtet man die vergleichsweise späte Entwicklung primitiv-mathematischer Ansätze (erste kombinatorische Erfolge in de Vetula um ca. 1250) oder gar die späte axiomatische Grundlage durch Kolmogorov (1933, die schon auf die gesamte Maßtheorie aufsetzen konnte), so kann man verstehen, dass auch die individuelle Entwicklung in Sachen Wahrscheinlichkeit gegenüber anderen mathematischen Begriffen verzögert, schwach ausgeprägt und störanfällig ist. In den Curricula hat man sich ehedem auf die Kombinatorik, also auf relativ soliden Grund, zurückgezogen. Nach einer Zwischenphase, in der man sich im Sog der Neuen Mathematik stark auf die mathematische (maßtheoretische) Struktur von Wahrscheinlichkeit bezog, hat man sich, eingeleitet durch die Anwendungswelle in den Curricula – zumindest fachdidaktisch – recht intensiv mit Methoden der Beurteilenden Statistik befasst. In den letzten zehn Jahren fokussiert man aber auf die Beschreibende Statistik – modern als „data handling“ gehandelt (siehe dazu auch Borovcnik, 2011). Wahrscheinlichkeit und statistische Methoden reduziert man dabei auf die Simulation von Daten (aufgrund von Hypothesen, was man leicht aus den Augen verliert) und deren Analyse. Die Chancen des Ansatzes liegen auf der Hand: statt aus hypothetischen Verteilungsannahmen und theoretischen Eigenschaften (mathematisch) Folgerungen zu ziehen, kann man sich auf die konkreten, simulierten Daten beschränken. Allerdings besteht auch die Gefahr, dass man den Wahrscheinlichkeitsbegriff völlig auf die frequentistische Deutung reduziert. Im vorliegenden Aufsatz versuchen wir, dem entgegenzusteuern und das Potential probabilistischer Begriffs- und Modellbildung zur Entfaltung zu bringen.

2. Brechen von Spaghettis – Relative Häufigkeiten und Modellieren

Relative Häufigkeiten werden schon sehr früh in der Wahrscheinlichkeitsrechnung in einen engen Zusammenhang mit Wahrscheinlichkeiten gebracht: Je mehr Daten, desto enger die Bindung und desto geringer der Unterschied zwischen beiden. Im Bernoullischen Gesetz der großen Zahlen wird der mathematische Hintergrund zur Rechtfertigung der Deutung von Wahrscheinlichkeit als relative Häufigkeit erbracht. Die stillschweigende Annahme dabei lautet, dass die Daten einem *Zufall*sexperiment entstammen; in der Beurteilenden Statistik sagt man, die Daten stammen aus einer Zufallsstichprobe.

Mit dem im Folgenden beschriebenen Experiment zum Brechen von Spaghettis in drei Teile (siehe auch Kataoka et al, 2009) kann man die grundlegende Bedeutung dieser Annahme besonders unterstreichen. Das Experiment endet üblicherweise damit, dass sehr viele Teilnehmer aus den drei

Teilen ein Dreieck bilden können. Die beobachteten„Erfolgsraten" liegen meist einiges über 65% und weichen damit sehr weit von der Wahrscheinlichkeit (0,25 im ersten Modell) ab, dass ein Dreieck gebildet werden kann. Im Vortrag haben 40 Personen mitgemacht und je zwei Spaghettis zerbrochen (die Teile wurden auf zwei getrennten Haufen geschichtet). Aus den 80 Haufen von 3 Teilen konnten sie insgesamt 61 Dreiecke bilden, das entspricht einer Erfolgsrate von etwas mehr als 76%. Wir müssen einsehen, dass wir die Spaghettis keineswegs zufällig brechen. Im Weiteren kann man den entstandenen Konflikt dazu nützen, die Art, wie wir tatsächlich die Spaghettis brechen, zu modellieren.

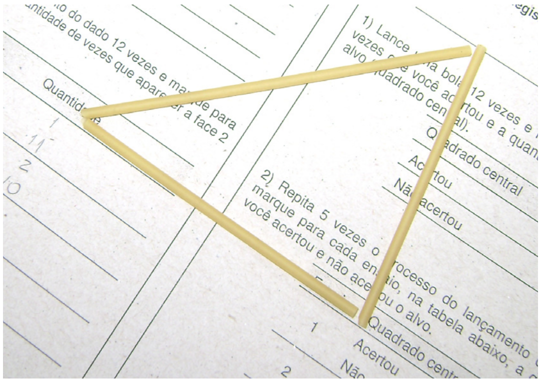

Abb. 1a: Die drei Teile lassen ein Dreieck bilden.

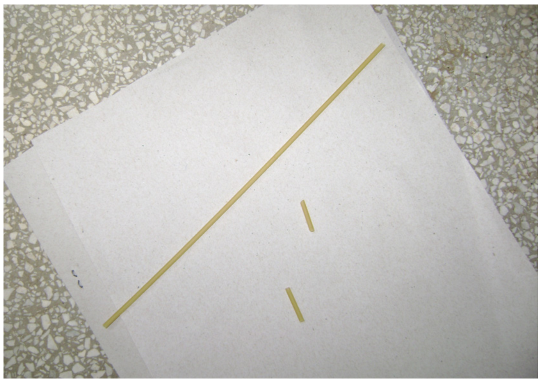

Abb. 1b: Hier kann kein Dreieck gebildet werden. [1]

Modelliert man das Brechen durch zwei Zufallszahlen aus $[0,1]$, so kann man die mathematische Lösung recht einfach erbringen. Unser Ziel war es jedoch, dieses zufällige Auswählen gleich zu simulieren und damit die Lösung zu schätzen. Erst werden die beiden Zufallszahlen z_1 und z_2 gewählt, dann bildet man das erste Bruchstück von 0 bis $\min\{z_1, z_2\}$, das zweite von $\min\{z_1, z_2\}$ bis $\max\{z_1, z_2\}$ und das dritte vom Rest auf 1. Anschließend prüfen wir, ob die Längen dieser Teile alle Dreiecksungleichungen einhalten (d.h., ob man ein Dreieck daraus bilden kann).

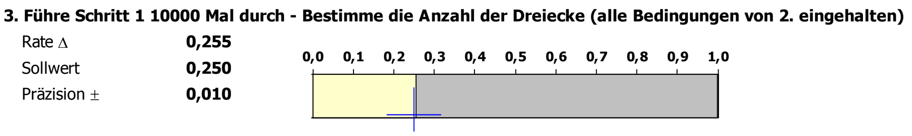

Abb. 2: Ausschnitt aus dem Tabellenblatt zum Ergebnis der Simulation von Modell 1.

Bei 10000 Simulationen erhalten wir eine Erfolgsrate von 0,255, die auch durch Wiederholen des gesamten Experiments (was in EXCEL ganz einfach per Knopfdruck möglich ist) weniger als ± 1% schwankt. Die Simulation schätzt die unbekannte Wahrscheinlichkeit recht genau. In gleicher Weise orientieren wir uns über die Schwankungsbreite bei 80 Daten und erhalten einen Schwankungsbereich von ca. ±10%. Mit der Erfolgsrate von 0,76 in unserem Experiment sind wir davon weit entfernt.

Ausgehend von der Überlegung, dass die meisten Menschen nicht wirklich zufällig brechen, sondern nach einem speziellen Schema vorgehen, passen wir unser Modell an. Wir brechen etwa die Nudel in zwei Stufen; zuerst brechen wir einmal, dann nehmen wir einen Teil und brechen diesen noch einmal. Wir beobachten auch, dass die meisten den größeren Teil ergreifen und brechen, nur wenige nehmen den kleineren. Die mathematische Lösung ist nun etwas schwieriger, wir umgehen diese aber mittels Simulation. Dazu wählen wir wieder zwei Zufallszahlen z_1 (für die erste Stufe) und z_2 (für die zweite Stufe). Gemäß der ersten Zufallszahl wird die Länge 1 in zwei Teile mit Längen z_1 und $1-z_1$ zerlegt. Das erste Teilstück hat demnach die Länge $\min\{z_1, 1-z_1\}$; das größere mit der Länge $\max\{z_1, 1-z_1\}$

[1] Die Fotos sind von Veronica Kataoka aus dem genannten Beitrag. Die Autorin hat sie uns freundlicherweise zur Verfügung gestellt.

wird gemäß dem Faktor z_2 geteilt. Wieder prüfen wir, ob alle Dreiecksungleichungen eingehalten werden.

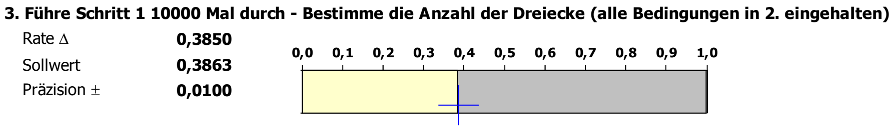

Abb. 3: Ausschnitt aus dem Tabellenblatt zum Ergebnis der Simulation von Modell 2.

Bei 10000 Simulationen erhalten wir eine Erfolgsrate von 38,5% und Wiederholungen des ganzen Experiments streuen wenig (wieder unter 1%). Allerdings sind wir noch immer weit davon entfernt, damit ein Modell zu haben, das ausreichend gut beschreibt, wie wir Spaghettis brechen. Die beobachtete Erfolgsrate von 76% liegt nämlich ganz weit weg von diesem Wert.

In einem weiteren Schritt versuchen wir, das Modell zu verbessern. Wir beobachten, dass nur wenige Personen zu kleine Stücke abbrechen. Entsprechend führen wir in das Modell einen Parameter q ein, welcher angibt, wie groß der Anteil der Bruchstücke mindestens sein muss. Wir müssen unsere Zufallszahlen nun aus dem Intervall $[q, 1-q]$ auswählen, d.h., $z_i = (1-2q)\cdot$ Zufallszahl() $+ q$.

Setzen wir in unserer Simulation $q = 1/10$, so können wir unsere Erfolgsrate für Dreiecke auf 56% erhöhen. Noch immer zu weit entfernt vom Ergebnis unserer Gruppe, aber schon weit besser. Wir erhalten noch größere Raten für Dreiecke mit größeren Werten für den Parameter q. Es scheinen aber noch weitere, nicht erfasste Eigenheiten, wie Personen Spaghettis brechen, eine Rolle zu spielen.

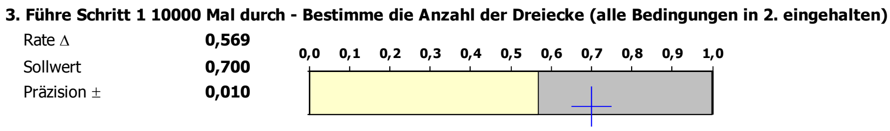

Abb. 4: Ausschnitt aus dem Tabellenblatt zum Ergebnis der Simulation von Modell 3 mit q = 1/10.

Damit Wahrscheinlichkeiten aus simulierten Experimenten hinreichend genau geschätzt werden können, muss man die erforderliche Zahl der Versuche erst bestimmen. Es gibt dazu einen ganz einfachen Weg: Zuerst nehmen wir eine Zahl, simulieren entsprechend viele Daten und schätzen daraus die Wahrscheinlichkeit. Dann wird die ganze Serie mit dieser Anzahl von Daten – auf Knopfdruck –einige Male wiederholt. Wie sehr nun die Schätzwerte schwanken, zeigt, wie genau wir schätzen können. Ist die Genauigkeit zu gering, dann erhöhen wir die Zahl der Daten und wiederholen das ganze Szenario bis die Schätzwerte nur mehr so wenig schwanken, dass wir diesen Fehler tolerieren können.

Man kann aber die mathematische Lösung informell ansprechen; diese ergibt sich aus der Länge von Vertrauensintervallen. Ein ($1-\alpha$)-Vertrauensintervall für eine unbekannte Wahrscheinlichkeit p lautet: $\hat{p} \mp z_{1-\alpha/2} \frac{\sqrt{\hat{p}\cdot(1-\hat{p})}}{\sqrt{n}}$ und da der Schätzwert $\hat{p}$ erst nach dem Experiment bekannt ist, kann man die (großzügige) Abschätzung $\sqrt{\hat{p}\cdot(1-\hat{p})} \leq \frac{1}{2}$ verwenden. Man kommt zum Schluss, dass man aus n Daten (aus einer Stichprobe) eine Wahrscheinlichkeit p auf $\mp\frac{1}{\sqrt{n}}$ genau schätzen kann (bei $1-\alpha = 0{,}95$). Solche überschlagsmäßigen Überlegungen kann man auch noch ohne theoretische Begründung anführen (man kann sie aber auch wiederum durch Simulationen plausibel machen). Auf diese Weise kann man eine Wahrscheinlichkeit auf ± 3% genau schätzen, wenn man 1000 Daten hat. Damit man also aus Simulationsszenarien genauer schätzen kann, benötigt man wohl etwa 10000 Daten, was eine Genauigkeit der Schätzung von ± 1% bewirkt. Das erklärt auch, warum sich Meinungsforschungsinstitute für nicht so wichtige Untersuchungen mit 500 Daten begnügen.

3. Zentrale Sätze und Eigenschaften von Stichproben

Es gibt zwei mathematische Satzgruppen, welche den Zusammenhang zwischen Daten aus Stichproben und Parametern der zugrunde liegenden Verteilung beschreiben. Die erste Gruppe betrifft die Gesetze der großen Zahlen, welche zeigen, wie und unter welchen Voraussetzungen die relativen Häufigkeiten (als Zufallsvariable) stochastisch gegen eine (unbekannte) Wahrscheinlichkeit konvergieren. In gleicher Weise konvergieren Mittelwerte stochastisch gegen den Erwartungswert einer Verteilung. Vorauszusetzen ist dabei, dass die Daten eine stochastisch unabhängige Folge von Zufallsvariablen (mit der Ausgangsverteilung) bilden; das und nichts anderes macht eine Stichprobe aus. Wenn Daten unter anderen Umständen gewonnen werden, dann wird es in der Regel schwierig, aus ihnen irgendwelche Parameter oder Wahrscheinlichkeiten zu schätzen. Das Kernproblem vieler Anwendungen ist, inwieweit die Bedingungen einer Stichprobe erfüllt sind. Hier sollen die Folgerungen aus dem *Vorliegen* einer Stichprobe nachvollzogen werden.

Weil die Mathematik doch schwieriger wird und die mathematische Abhandlung die Eigenschaften sehr abstrakt erscheinen lässt, bietet es sich hier an, die Konsequenzen durchzuspielen, also über Simulationen erlebbar zu machen. Grenzprozesse bleiben jedoch endlichen Simulationen verschlossen; außerdem sind sie damit schwer zu verstehen. Daher variieren wir die Länge der Serie (den Umfang der Stichprobe) systematisch und können so den resultierenden Effekt aufzeigen. Die Streuung der relativen Häufigkeiten wird, bei zunehmendem Umfang, kleiner. Als gedankliche Extrapolation schließt sich daran an, dass die Streuung gegen Null konvergiert, wenn der Stichprobenumfang über alle Maßen wächst.

Die zweite Satzgruppe betrifft die Zentralen Grenzverteilungssätze, die aufzeigen, dass – unter geeigneter Standardisierung – die Verteilung der relativen Häufigkeiten bzw. der Mittelwerte (jeweils als Zufallsvariable) gegen die Normalverteilung konvergiert. Hier belassen wir es bei Verweisen auf andere Quellen. Die mathematische Herleitung wird noch komplizierter und wird auch auf Universitätsniveau allenfalls mit Lücken angeboten. Die Methode der Veranschaulichung über Simulation wird auch daher noch wesentlicher. Sie wird aber auch aus einem anderen Grund existentiell: Wenn etwa Mittelwerte (als Zufallsvariable) stochastisch gegen einen Punkt (nämlich den Erwartungswert) konvergieren, wie kann man dann noch von einer Verteilung des Mittelwerts sprechen? Durch die Umstände einer Simulation wird augenfällig, wie man das völlige Zusammenziehen der Verteilung durch Standardisierung verhindern kann.

Wenn Stichproben (die unabhängige Wiederholung desselben stochastischen Experiments) erst die Konvergenz zu den interessierenden Parametern ermöglichen, sind Stichproben *die* Voraussetzung, dass man diese Parameter schätzt. Es verbleibt das Paradoxon, dass rein zufälliges Auswählen aus einer Grundgesamtheit die beste Methode liefern soll, um über die Grundgesamtheit etwas in Erfahrung zu bringen. Dabei steht zufälliges Auswählen synonym dafür, dass man alle Kenntnis über die Elemente einer Grundgesamtheit bewusst ausblendet. Dem steht gegenüber, dass etwa Meinungsforschungsinstitute über so genannte Quotenstichproben ihre Information gewinnen und die interessierenden Merkmale schätzen. Ihr Erfolg zeigt, dass der Weg, vorhandene Information über die Elemente der Grundgesamtheit auszunützen (um die bekannten Quoten meist soziographischer Merkmale zu erfüllen) so schlecht nicht sein kann. Im Abschnitt Eigenschaften von Stichproben stellen wir einen Kompromiss von Quoten und zufälligem Auswählen der Elemente der Stichprobe vor: das geschichtete zufällige Auswählen verbessert die Präzision der Schätzungen aufgrund von Stichproben enorm. Es kann also von Vorteil sein, Kenntnisse über die Grundgesamtheit in die Rekrutierung von Stichproben einzubringen.

Gesetze der großen Zahlen

Vom Verhältnis zwischen relativen Häufigkeiten und Wahrscheinlichkeiten war schon die Rede. Grundlegend zum besseren Verständnis der Zusammenhänge ist die Schwankungsbreite des Zufalls und wie sich diese verringert, wenn man mehr (zufällige) Daten hat. Ein einfaches Experiment dazu ist die Auswahl von Zufallsziffern aus der Menge der Ziffern 0 bis 9. Es geht in diesem Experiment zum einen um die Art des Einflusses (Monotonie) der Zahl der Daten. Zum anderen ist es wichtig, ein Gefühl zu entwickeln, wie stark der Zufall bei bestimmten Referenz-Anzahlen von Daten schwankt.

Wir erzeugen 50 – gleich verteilte – Zufallszahlen und werten die relativen Häufigkeiten aller Ziffern aus. Ein Stabdiagramm der relativen Häufigkeiten wird zur Animation der Schwankungen, wenn man die Zufallszahlen dynamisch neu auswählt. Die Stabdiagramme werden wackeln, insgesamt wird aber kaum eine der Staffeln aus einem Bereich von ± 10% herausfallen. Das Band der Schwankungen ist in Abb. 5 leicht unterlegt. Wiederholt man das gesamte Szenario mit 1000 statt der 50 Zufallszahlen, so ist es einmal offensichtlich, dass das Band der Schwankungen sich stark eingeschnürt hat. Mit 1000 Zufallszahlen kann man die zugrunde liegende Wahrscheinlichkeit von 1/10 für jede Ziffer viel genauer schätzen als mit 50. Wiederholt man die Simulation öfter, so kann man erkennen, dass die Schwankungsbreite in etwa bei ± 2% liegt.

Das Band der Schwankungen ist auf eine spezielle Zahl (wir könnten uns nur auf die 7 konzentrieren) abgestimmt. Es wird daher bei allen Zahlen doch in etwa einem Viertel der Versuche wenigstens eine Staffel außerhalb des Bandes liegen. Hier kann man auf den Unterschied zwischen i) Warten, dass die 7 außerhalb des normalen Bandes der Schwankungen liegt, und ii) Warten, dass irgendeine der Zahlen außerhalb liegt, hinweisen. Dies wird später beim statistischen Testen relevant, wenn man erst die Daten auswertet, dann schaut, welche Werte extrem sind und dann nur diese testet. Eigentlich muss man schon vorher ausmachen, auf welche Merkmale man schaut (wie sich die 7 verhält) und nicht erst auf die Daten warten, bemerken, dass die 3 extrem stark vertreten ist, und dann behaupten, man hätte von Anfang an auf 3 geschaut. Wie man bei mehrfachen Tests die Irrtumswahrscheinlichkeiten korrigiert, liest man – ohne viel Mathematik – in Lorenz (1996) nach.

Üblicherweise untersucht man die „Konvergenz" der relativen Häufigkeiten $h_n = \frac{s_n}{n}$ in Abhängigkeit von der Länge eines Versuchs anhand einer (einzigen) Serie von Versuchen. Man zeichnet in das Diagramm den Verlauf der Punkte (n, h_n) und sieht, wie die relativen Häufigkeiten gegen (ja gegen welchen Wert eigentlich?) „konvergieren". Und in Wirklichkeit konvergieren sie ja nicht, es handelt sich ja keineswegs um eine Folge im Sinne der Analysis.

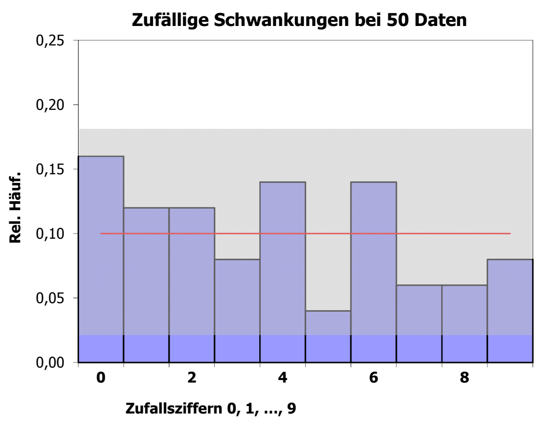

Abb. 5a: Bandbreite der Schwankungen ist breit: +/–0.08.

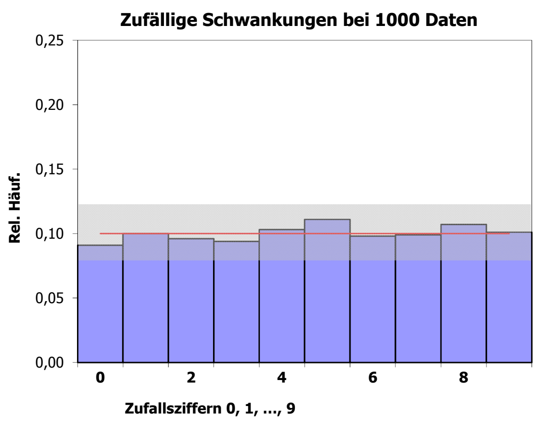

Abb. 5b: Bandbreite ist schmal: 0.02.

Das Bernoulli-Gesetz der großen Zahlen spricht ja auch eine andere Sprache: Ist $\frac{S_n}{n}$ die relative Häufigkeit von 1 in einer (stochastisch) unabhängigen Wiederholung eines Versuches mit den zwei Ausgängen 0 und 1 und ist bei jedem Versuch die Wahrscheinlichkeit für 1 gleich p, so gilt:

$P(|\frac{S_n}{n} - p| \geq \varepsilon) \to 0$, falls $n \to \infty$ ($\varepsilon > 0$ darf beliebig gewählt werden).

Es treten in diesem Gesetz nirgends Zahlenfolgen auf, deren Konvergenz man zu prüfen hat. (Das wäre erst im starken Gesetz der großen Zahlen der Fall, das aber jenseits von Schule liegt.) Es geht um Zufallsvariable, deren Verteilung mit zunehmendem n stärker *innerhalb* des Fensters $[p-\varepsilon, p+\varepsilon]$ fällt. Man kann daraus auch Fenster mit abnehmender Breite ableiten, deren Wahrscheinlichkeiten aber dann gleich sind. Letzteres haben wir in unserem Experiment veranschaulicht: das Zufallsband für die relativen Häufigkeiten wird immer enger, wenn die Zahl der Daten zunimmt.

Bernoulli-Experimente sind Prototypen für Zufallsexperimente. Wir haben zwei Versuchsausgänge (mit 0 und 1 codiert; Ereignis E ist ausgeblieben bzw. eingetreten). Die Wahrscheinlichkeit für 1 ist dabei p. Wiederholt man das Experiment (stochastisch) unabhängig voneinander, so bekommt man mit der Summe der Daten die absolute Häufigkeit der 1en und damit des Auftretens des Ereignisses E. Bestimmt man die relative Häufigkeit von E, so erhält man damit einen Schätzwert für die Wahrscheinlichkeit $P(E)$.

Mit diesem wechselseitigen Verhältnis von relativer Häufigkeit und Wahrscheinlichkeit ist eine Reihe von Vorstellungen zu kalibrieren. Man kann sich dabei an Bernoullis Gesetz der großen Zahlen orientieren. Wie schon gesagt, die Konvergenz betrifft eigentlich die *Verteilung* (ein Zusammenziehen der Verteilung mit zunehmender Anzahl von Daten). Man stellt sich jedoch intuitiv einfacher die Folgen der relativen Häufigkeiten vor, wie diese „konvergieren". Die Animation (siehe Abb. 5a und b) sollte den Fokus auf die Verteilung und die Schwankungsbreite des Zufalls lenken.

Die Konvergenz der Folgen der relativen Häufigkeiten wird jedoch (intuitiv) gerne (zu Unrecht) auf eine Konvergenz der absoluten Häufigkeiten übertragen. In einigen Tiefeninterviews mit Kindern zu dieser Konvergenz werden Experimente mit Summen (absoluten Häufigkeiten) durchgespielt. Im Klassengespräch oder im Gespräch Lehrer–Kinder ist von Konvergenz die Rede (siehe etwa Schnell, 2011). Das kann nur verwirren und die Antworten der Kinder kann man eigentlich nicht mehr interpretieren. Es zeigt sich, dass es wichtig ist, das Verhalten der absoluten und relativen Häufigkeiten streng auseinander zu halten: die relativen Häufigkeiten konvergieren (im Sinne der Verteilung), die absoluten Häufigkeiten aber divergieren. Je länger die Serie von Zufallsversuchen anhält, umso stärker streuen die Ergebnisse der absoluten Häufigkeiten: Bei 100 Versuchen ±10, bei 1000 Versuchen ±30 (95% Sicherheit).

Divergenz der absoluten Häufigkeiten (Pendant zum Bernoulli-Gesetz der großen Zahlen):

$P(|S_n - np| \geq w) \to 1$, falls $n \to \infty$ ($w > 0$ darf dabei beliebig groß gewählt werden).

Wir stellen zwei Experimente zur Klärung der intuitiven Vorstellungen vor. Diesmal werden wir nicht simulieren, sondern die Kenntnis der Binomialverteilung ausnutzen (oder einen Vorgriff darauf machen und gleichzeitig den Weg zu dieser Verteilung vorbereiten). Die absolute Häufigkeit S_n ist für das unabhängig wiederholte Bernoulli-Experiment binomial verteilt mit den Parametern n und p.

In einer Tabellenkalkulation sind die einzelnen Wahrscheinlichkeiten schnell abgerufen; Abb. 6 zeigt einige ausgewählte Verteilungen. Dabei ist das Fenster um den Erwartungswert np mit der Breite $\pm w$ hell, der Rest hingegen zur Unterscheidung kräftig eingefärbt. Für n = 50 liegt die gesamte Wahrscheinlichkeit innerhalb des Fensters, bei n = 100 liegt schon einiges außerhalb, bei 500 noch mehr und bei 1000 liegt nur mehr wenig innerhalb. Wir können die Erfolgswahrscheinlichkeit p jetzt mit einem Schieberegler ändern und sehen, die Wahrscheinlichkeit, außerhalb des Fensters fester Breite zu liegen, nimmt mit zunehmendem n immer zu und das gilt unabhängig vom tatsächlichen Wert von p.

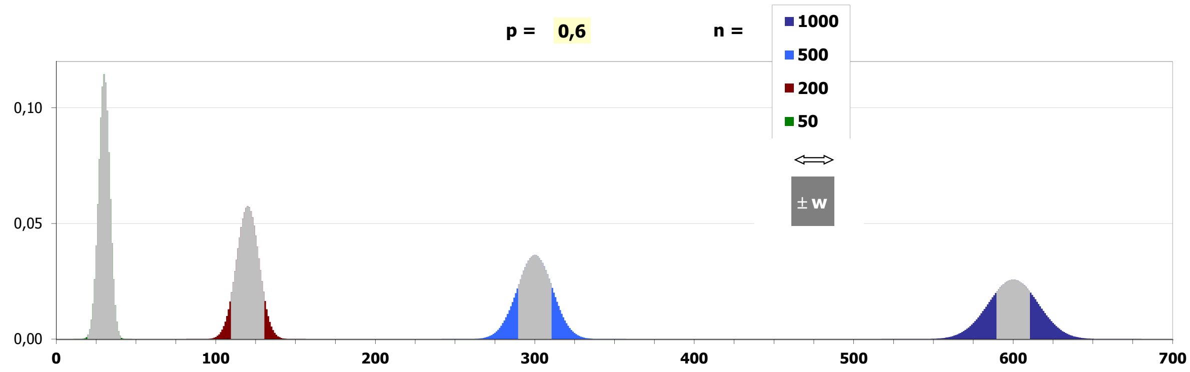

Abb. 6: Divergenz der absoluten Häufigkeiten: Wenn die Zahl der Experimente zunimmt, verbleibt mehr und mehr Wahrscheinlichkeit außerhalb des Fensters mit fester Breite ± w.

Die relativen Häufigkeiten $H_n = \frac{S_n}{n}$ entstehen durch Dividieren mit dem Stichprobenumfang, das ändert die Wahrscheinlichkeiten selbst nicht; anschaulich gesprochen wird das Stabdiagramm einfach gestaucht und verschoben. Wir untersuchen nun – wie eben – wie viel Wahrscheinlichkeit sich außerhalb eines Fensters fester Breite $\pm\varepsilon$ um den Erwartungswert p befindet. Wir sehen in Abb. 7, dass jetzt mit zunehmendem Umfang n immer weniger Wahrscheinlichkeit außerhalb dieses Fensters liegt.

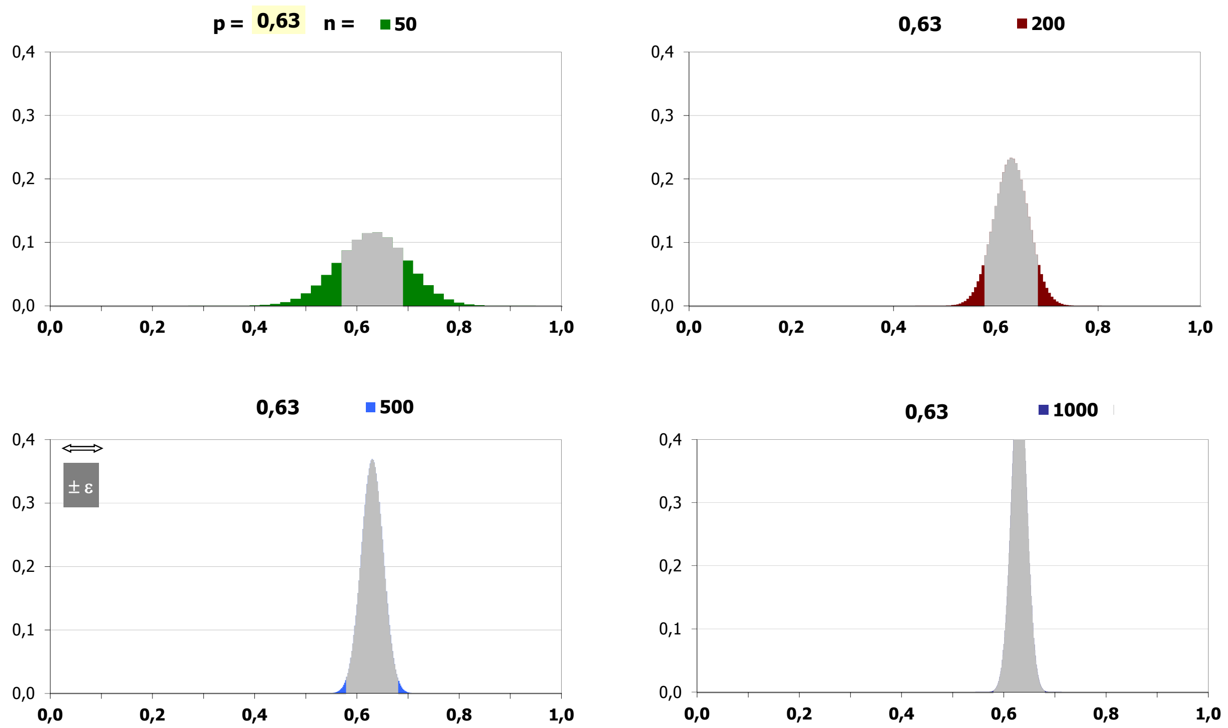

Abb. 7: Konvergenz der relativen Häufigkeiten: Wenn die Zahl der Experimente zunimmt, verbleibt mehr und mehr Wahrscheinlichkeit innerhalb des Fensters mit fester Breite ± ε .

In diesem Beispiel ist die Animation einer Simulation vorzuziehen. Die zugrunde liegende Binomialverteilung wird dabei nicht durch Simulation rekonstruiert sondern direkt verwendet. Dann werden ihre Parameter verändert, sodass man die Auswirkungen dynamisch – wie in einem Film – verfolgen kann. Zum einen wird die Binomialverteilung (zumindest) als Arbeitsprogramm eingeführt (wenn sie nicht schon behandelt wurde), zum anderen gibt es beim Zufall in der Simulation immer wieder irreguläre Ergebnisse, die nicht verallgemeinerbar sind, womit man den eigentlichen Effekt nicht mehr so deutlich erkennen kann. Um dies hier zu vermeiden, bräuchte man sehr viele gleichzeitig ablaufende Simulationen und das würde EXCEL an seine (Leistungs-)Grenzen führen.

Eigenschaften von Stichproben

In diesem Abschnitt zeigen wir, wie man die Präzision der Schätzung von Parametern aufgrund von Stichproben enorm verbessern kann, wenn man Kenntnisse über die Struktur einer Grundgesamtheit einbringt und zufälliges Auswählen auf Schichten beschränkt, welche nach Vorkenntnissen eine kleine Streuung im Hinblick auf das Zielmerkmal aufweisen. Ein extremes Beispiel wäre etwa eine Schichtung der Grundgesamtheit in zwei Untergruppen, in denen das Zielmerkmal konstant ist. Weiß man das, so reicht eine Stichprobe vom Umfang 2, um den Wert exakt zu kennen, während man mit zufälligen Daten auch aus großen Stichproben immer noch mit einer großen Streuung (Ungenauigkeit) zu rechnen hätte. Wir betrachten eine in zwei homogene Schichten zerfallende Grundgesamtheit und wollen den Mittelwert aus einer Stichprobe schätzen.

In Anlehnung an den Verbraucherpreisindex (VPI) erheben wir einen Index von Hemden bestimmter Qualität. Das aktuelle Preisniveau wird durch eine Stichprobe (beim echten VPI ist das keine Zufallsstichprobe) erhoben. Die Grundgesamtheit bestehe aus 30 Geschäften, davon 10 in der normalen, 20 in der gehobenen Klasse.

Wir verwenden zwei verschiedene Methoden zur Ziehung der Stichprobe:

i) einfache Zufallsstichprobe von 6 aus allen;
ii) geschichtete Zufallsstichprobe von 2 aus der Schicht der billigen und 4 aus der Schicht der teuren Geschäfte.

Die Größenordnung des Fehlers, der bei der Schätzung des echten Preisniveaus (für alle Geschäfte) aus der Stichprobe eintreten kann, wird durch wiederholtes Ziehen von Stichproben nach dem jeweiligen Schema augenfällig.

Während für 1000 Stichproben die Schätzwerte aus den einfachen Zufallsstichproben zwischen 3 und 12 schwanken (Abb.8 Mitte), ist die Fluktuation aus der geschichteten Stichprobe in der Größenordnung von ±1 (unten). Bei Wiederholung des gesamten Szenarios erhält man ein ähnliches Bild. Die gewöhnliche Zufallsstichprobe liefert eine weitaus schlechtere Schätzung als die geschichtete.

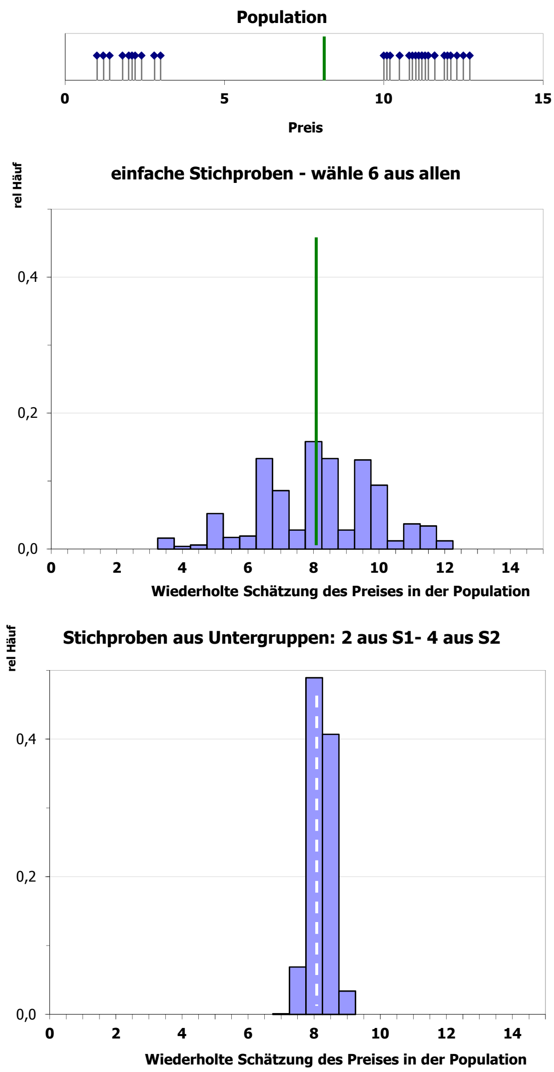

Abb. 8: Wiederholte Schätzung des Preisniveaus bei einfachen und bei geschichteten Zufallsstichproben.

Zuerst hatten wir die Bedeutung von Zufallsstichproben belegt, damit die Daten relevante Aussagen über die Grundgesamtheit erlauben. Jetzt sehen wir, dass man in vielen Fällen „besser als Zufall" werden kann, wenn man Information über die Grundgesamtheit einbringt: Das Merkmal Preis von Hemden streut sehr wenig innerhalb der jeweiligen Schicht. Diese Kenntnis kann man ausnutzen und das Schätzverfahren viel präziser machen. Zufälliges Auswählen ist gleich zu setzen mit „keine Kenntnis über die Grundgesamtheit – alle Elemente haben dieselbe Chance ausgewählt zu werden", die Quotenmethode bedeutet, dass gewisse Quoten in der Stichprobe erfüllt sein müssen (etwa Geschäfte in urbanen bzw. in ländlichen Bereichen). Geschichtete Stichproben orientieren sich an Schichten der Population, welche hinsichtlich des Zielmerkmals wenig streuen. Bringt man diese Kenntnis mit ein und verbindet die Auswahl mit Zufall innerhalb der Schichten, so wird die Schätzung extrem verbessert. In der Praxis allerdings kennt man oft wohl relevante Schichten, kann aber, mangels Zugänglichkeit nicht zufällig aus ihnen auswählen.

Zentraler Grenzverteilungssatz und Approximation von Verteilungen
Viele Verteilungen sind untereinander approximierbar. Die hypergeometrische (entsteht beim Ziehen ohne Zurücklegen aus einer Urne mit zwei Arten von Kugeln) lässt sich durch die Binomialverteilung (Ziehen mit Zurücklegen) approximieren. Aus der Modellsituation des Ziehens aus einer Urne erkennt man leicht, dass es wenig Unterschied macht, wenn der Auswahlsatz (die Anzahl gezogener Kugeln bezogen auf die Anzahl aller Kugeln in der Urne) klein ist.

Schwieriger wird die Situation schon, wenn man die Approximation der Binomial- durch die Poisson- oder durch die Normalverteilung rechtfertigen will. Dahinter stehen jeweils Grenzwertsätze. Auch diese kann man durch Simulationen und Animationen motivieren. Die Mathematik wird auch auf Universitätsniveau vielfach umgangen. Es gilt also, die Zusammenhänge plausibel zu machen. Als ein Nebenergebnis ergibt sich eine wichtige Deutung für die Normalverteilung:

Immer wenn man (fiktiv) ein Merkmal X in eine Summe von kleinen Summanden zerlegen kann, welche unabhängige und identisch verteilte Zufallsvariable sind, wird man die Verteilung von X näherungsweise durch eine Normalverteilung beschreiben können. Die Szenarien sind in Borovcnik und Kapadia (2011) beschrieben; in Borovcnik (o.J.) findet man eine Realisierung in EXCEL.

Die Binomialverteilung ergibt sich auch als Summe von 0,1-Merkmalen (wiederholtes Drehen am selben Glücksrad mit 2 Sektoren, oder Ziehen aus einer Urne mit zwei Arten von Kugeln – jeweils wird die Summe der 1en gezählt). Hat man diese Struktur der Binomialverteilung einmal hervorgehoben, so sollte die Möglichkeit der Approximation durch eine Normalverteilung nicht überraschend sein. Durch Animationen oder Simulationen kann man neben dem Umstand, dass sich für eine größere Anzahl von Summanden näherungsweise eine Normalverteilung ergibt, auch die Einsicht ermöglichen, dass die Konvergenzgeschwindigkeit zur Normalverteilung ganz wesentlich von der Gestalt der Verteilung der Summanden abhängt: Ist diese symmetrisch und hat selbst schon viele Werte, so reichen schon fast 3 Summanden aus (die Augensumme von drei Würfeln etwa), ist diese sehr schief und hat wenige Werte, so sind 100 nicht genug.

4. Stochastisches Modellieren

Der Kontext einer Aufgabe wird häufig als Feigenblatt verwendet; man könnte die Geschichten fast beliebig austauschen, es geht eigentlich fast immer um das Einüben einfacher Fertigkeiten. Für eine solche Kritik siehe Eichler und Vogel (2010). In vielen Aufgaben soll man Wahrscheinlichkeiten berechnen; eigentlich muss man nur die richtige Verteilung auswählen, der Rest ist Routine. Selten gilt es, eine Frage damit zu beantworten, oder gar alternative Modelle in Erwägung zu ziehen oder die Gültigkeit der Annahmen zu diskutieren. Man erhält und braucht kaum eine klare Vorstellung davon, wie man sich durch probabilistisches Modellieren in der realen Situation im Kontext der Aufgabe einen Vorteil verschafft. Das volle Potential von Wahrscheinlichkeit zur Modellierung wird in einem solchen Ansatz verpasst (vgl. Borovcnik & Kapadia, 2011).

Zum einen wollen wir mit den Parametern der verwendeten Verteilung spielen, um die Auswirkungen auf die Lösung zu studieren und gleichzeitig uns damit beschäftigen, dass die Voraussetzungen oft auch gar nicht so gut zutreffen. Zum anderen wollen wir den Zufall als Modell gegen andere – innere Erklärungen – von Phänomenen ausspielen. Dabei werden innere Zusammenhänge als „Ursachen" vom Kontext her verstanden, wogegen der Zufall immer eine Anbindung an den Umstand, dass überhaupt keine Zusammenhänge bestehen, erlaubt. Schon die zufällige Auswahl von nummerierten Kugeln aus einer Urne (wie beim Lotto) soll eben keine Bevorzugung irgendeiner Kugel repräsentieren. Es gibt keinerlei Grund, dass irgendeine eher gewählt wird. Das wird den Spielern gegenüber als fair aufgefasst, keiner ist besser dran als ein anderer.

Parkplatzproblem – kann man mehr Tickets verkaufen als man Plätze hat?

Viele Luftfahrtgesellschaften nutzen den Umstand, dass ein kleiner Prozentsatz von verkauften Tickets nicht in Anspruch genommen wird, weil die Personen aus unterschiedlichsten Gründen nicht auftauchen. Warum dann nicht mehr Tickets verkaufen als man Plätze hat? An der Klagenfurter Universität hat man im Zuge der Einführung von Parkgebühren auch daran gedacht, mehr Parkkarten zu verkaufen als Parkplätze vorhanden sind. Das Beispiel hat also durchaus realen Hintergrund. In der üblichen Formulierung ergibt sich bei der vorgegebenen Zahl der Plätze, der Wahrscheinlichkeit, dass ein Ticketkäufer seinen Platz beansprucht, und der Zahl der verkauften Tickets eine Wahrscheinlichkeit für die Überbelegung: es kommen mehr als man Plätze hat. Ein Risiko, dessen Größe man zu interpretieren hat. Wozu man diese Zahl dann nützen kann, ist nicht mehr Thema der Aufgabe.

Durch Anpassen der Parameter mit den Schiebereglern kann beobachtet werden, wie sich die Verteilung ändert – siehe dazu Abb. 9. Man berechnet nicht nur die Wahrscheinlichkeit (das Risiko) einer Überbelegung, dass also mehr Leute kommen als Plätze vorhanden sind sondern sieht auch die einzelnen Möglichkeiten der Überbelegung im Detail (im Original sinnigerweise mit roten Balken dargestellt). Es wird auch klar, wie groß die einzelnen Überbelegungen werden können.

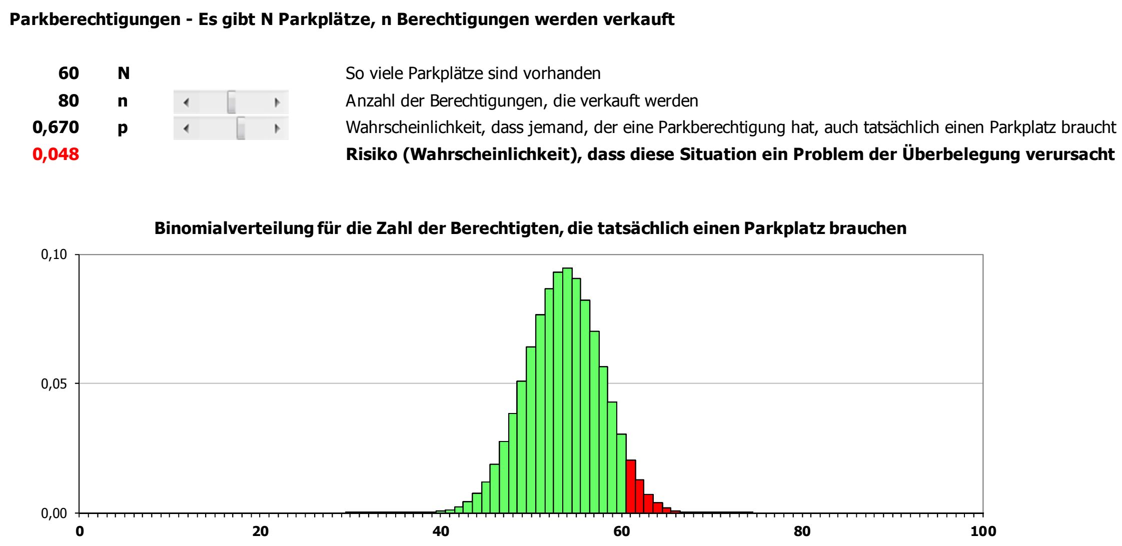

Abb. 9: Ausschnitt aus dem Tabellenblatt zum Parkplatzproblem.

Das ist im Hinblick auf Kollisionsstrategien wesentlich; was macht man, wenn sehr große Überbelegungen auftreten. Bei den Flugtickets werden die Fluggäste vertröstet, verköstigt (auch die Übernachtung wird bezahlt), es wird auch ein Erstattungsbetrag angeboten, wenn andere Leute auf ihren Platz verzichten. Die logistischen Probleme der Umbuchungen sind bei kleiner Überbuchung leicht lösbar, bei größeren Überbuchungen wird es schwierig; vor allem die Verlängerung der Reisezeit wird zu einem Problem, an dem die Akzeptanz der angebotenen Ersatzlösungen scheitern kann.

Modell: Jeder Studierende mit Ticket wird so modelliert als ob unabhängig von den anderen eine Münze mit Wahrscheinlichkeit p für Kopf geworfen wird, bei Kopf wird der Parkplatz beansprucht, bei Zahl eben nicht. Die Summe derer, die einen Parkplatz tatsächlich beanspruchen folgt dann einer Binomialverteilung mit n (Ticketinhaber) und p.

Modellkritik: Es ist wichtig, auf Gegebenheiten in der Realität hinzuweisen, die von den Modellbedingungen abweichen. Zunächst ist eine allen Studierenden gemeinsame Wahrscheinlichkeit p eine reine Fiktion, der man aber als Durchschnittsargument doch noch einiges abgewinnen kann. Allerdings wird p über die Zeit hinweg nicht konstant sein, etwa am Abend oder am späten Vormittag, oder über die Zeiten des Semesters hinweg. Dann aber geht es um die Unabhängigkeit der Studierenden, welche durch Umstände wie gemeinsame Großlehrveranstaltungen, große Prüfungstermine etc. gestört wird. Insgesamt kann man dem Modell zuschreiben, dass es nicht so gut passt, sondern eher im Sinne eines Szenarios (Borovcnik 2009a) zu verstehen ist: Obwohl es nicht übermäßig gut passt, kann man Anhaltspunkte über kritische Parameter, die Sensitivität der Lösung im Hinblick auf Veränderungen der Werte von Einflussparametern, und die Richtung von Verbesserungen in Zielgrößen bekommen.

Durch Anpassung von Tabellenblättern kann man nach kurzer Zeit eine vielseitige Bibliothek aufbauen. Das obige Blatt wurde leicht modifiziert, um eine andere Aufgabenstellung, die auch zur Binomialverteilung führt, zu bearbeiten. Es geht dabei zunächst um den so genannten *N*-Effekt. Die Psychologen Tor und Garcia haben aus ihren Experimenten folgendes Gesetz herausgefiltert: Je größer die Anzahl der Teilnehmer in einem Wettbewerb, desto kleiner die *Motivation* der Teilnehmer, sich anzustrengen. Interessant ist, dass neben der Erklärung durch einen inneren Zusammenhang (Motivation) auch eine Erklärung durch den reinen Zufall möglich ist. Der Zufall übernimmt hierbei die Rolle der völligen Abwesenheit von inneren Faktoren, also ganz ohne jeden Grund. Die stochastische Erklärung kann man mit einer Animation verständlich machen. Für Details der wissenschaftlichen Diskussion siehe König (2011).

Unabhängige Würfelexperimente zur Erklärung von Placebo-Effekt und der Regression zur Mitte

Das folgende Experiment mit 42 Würfeln wird von uns benützt, um zwei Phänomene zu erklären: den Placebo-Effekt und das Phänomen der Regression (des Zurückschreitens) zur Mitte. In beiden Fällen simulieren wir ein Zufallsexperiment zweimal, und zwar unabhängig voneinander. Wir werfen die Würfel das erste Mal und notieren die 1en als Flops und die 6en als Tops. Wir ziehen nur diese Würfel für den zweiten Wurf heran und beobachten, wie sich die Tops und Flops jetzt verhalten. In Abb. 10 sind die Ergebnisse des zweiten Wurfs zusammen gefasst: Die Flops erreichen (diesmal) im Durchschnitt 3,83, die Tops nur 3,11. Die Flops erweisen sich nun – dieses eine Mal – als besser.

Jetzt könnte man das als Zufall abtun. Wenn man jedoch das Szenario 1000mal wiederholt und auswertet, erkennt man, dass sich die Tops und Flops im 2. Wurf ganz ähnlich verhalten (in einer Serie waren in 50,8% die Flops besser als die Tops). Warum sollten sie denn Unterschiede zeigen, wir haben doch den zweiten Wurf als unabhängiges Experiment durchgeführt.

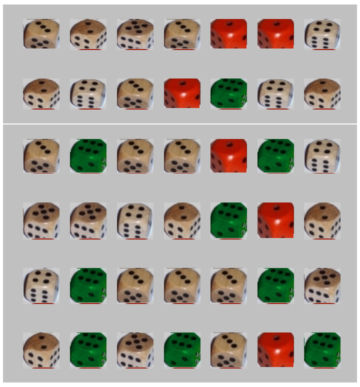

Abb. 10a: Ergebnis des 1. Wurfs von 42 Würfeln.

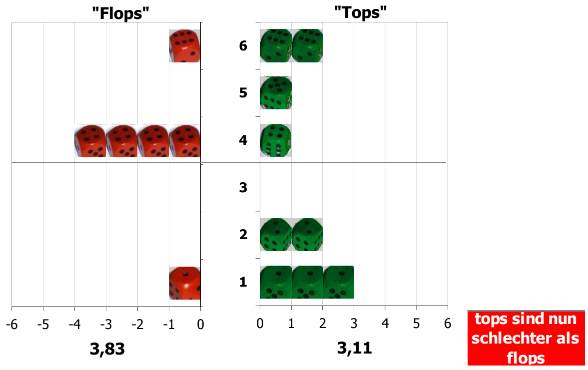

Abb. 10b: 2. Wurf – verglichen werden die besten (Tops) und die schlechtesten (Flops) des ersten Wurfs.

Während die Tops und Flops beim zweiten Wurf ähnlich abschneiden, ergibt sich in der Entwicklung zum ersten Wurf eine Tendenz:

- die Tops des 1. Wurfs werden schwächer,
- die Flops des 1. Wurfs werden besser.

Erklärung des Placebo-Effekts (nach Dubben & Beck-Bornholdt, 2007): Wir unterstellen, dass wir zum Arzt gehen, wenn wir uns schlecht fühlen. (entspricht dem Flop beim 1. Wurf) und fühlen uns danach besser, egal was wir bekommen (entspricht der Verbesserung im unabhängigen 2. Wurf). Die Verbesserung erklärt sich hier aus dem Vergleich mit unabhängigen Zufallsversuchen und nicht als Folge eines besonderen Placebo-Effekts.

Die unabhängige Wiederholung eines Zufallsexperiments hat mit dem vorausgehenden keinerlei sachlichen oder ursächlichen Zusammenhang; sie steht auch nicht mit anderen Einflüssen, die sich zwischen den beiden Experimenten ergeben, in irgendeiner Verbindung. Der Umstand, dass es jenen, denen es vorher (im ersten Wurf) schlecht ging, nachher im Durchschnitt wesentlich besser geht, wird hier durch die Veränderung der Ergebnisse in einem unabhängigen Zufallsversuch erklärt. Basis der Erklärung der Veränderung ist der reine Zufall in seiner unabhängigen Wiederholung und damit das Fehlen jeder kausalen Einwirkung.

Erklärung des Phänomens der Regression zur Mitte: Schon zu Beginn der Regressionsrechnung stand bei Karl Pearson der Regressionseffekt im Mittelpunkt (siehe Borovcnik 1988): Söhne großer Väter tendieren dazu weniger groß zu sein als ihre Väter; Söhne kleiner Väter tendieren dazu, größer zu sein als ihre Väter. Im Mittel sind die Söhne im ersteren Fall kleiner und näher zum Mittelwert der Population, die letzteren sind im Mittel größer: die Söhne tendieren zur Mitte. Die Methode hat nach diesem Phänomen ihren Namen erhalten. Das Phänomen der Regression zur Mitte ist jedoch ein reines Artefakt, d.h., sie ist rein durch Zufall erklärbar. Ohne auf die technischen Details der Regressionsrechnung hier einzugehen, kann man dies mit dem Würfelexperiment von oben einsehen:

Im ersten Wurf wird die Körpergröße der Väter ermittelt, im zweiten die der Söhne. Wir haben im Experiment nur die Tops (die größten Väter) und die Flops (die kleinsten Väter) weiter analysiert. Durch Zufall (und nicht durch Vererbung) haben wir die Körpergröße der Söhne ermittelt. Mit dem schon angesprochenen Ergebnis: Die Söhne der größten Väter werden im Durchschnitt kleiner, die Söhne der kleinsten Väter werden im Durchschnitt größer sein (mehr zum Mittelwert neigen).

Solche Effekte einer Regression zur Mitte findet man in Wetterphänomenen, in Sportbewerben, in denen – abgesehen von einer dünnen Spitze – ein großes Feld der Bewerber nahezu durch Zufall seinen Rang bei den einzelnen Bewerben einnimmt.

5. Beurteilende Statistik

In der Beurteilenden Statistik gibt es (mindestens) zwei Richtungen, denen die großen Verständnisprobleme zuzuordnen sind. Die eine besteht in der vorherrschenden Denkweise, Möglichkeiten, die unterstellt werden, miteinander zu vergleichen. Es ist in der Angewandten Mathematik (ohne Stochastik) schon schwierig, zwei (oder mehrere) verschiedene Modelle miteinander zu vergleichen: Warum werden diese zwei gewählt? Nach welchen Kriterien kann man diese vergleichen? Was ist, wenn andere Modelle gewählt würden? Warum werden überhaupt zwei gewählt, eines der Modelle muss doch besser sein?

In der Stochastik kommt noch hinzu, dass die Modelle Wahrscheinlichkeitsverteilungen sind. Deren Folgerungen sind mit fast allen realen Daten vereinbar, sprich, viele noch so abwegige Verteilungen haben eine von 0 verschiedene Wahrscheinlichkeit für das Auftreten von beobachteten Daten (Ereignissen). Die Kriterien werden zu Indizien. Es kommt hinzu, dass man aufgrund von Wahrscheinlichkeiten entscheidet, das eine oder das andere Modell zu favorisieren. Fast nie kann man sich davon frei machen, dass Fehler in den Entscheidungen möglich sind. Die Kennziffern, aufgrund derer man die Entscheidungen trifft, sind auch nur indirekt auf den vorliegenden Anwendungsfall zu beziehen und zeigen nur beschränkt, ob die Entscheidung gut ist: Fehler 1. Art, Fehler 2. Art, p-Wert (beobachtete Signifikanz), Überdeckungswahrscheinlichkeit.

Wir werden zum einen durch einfache Experimente den Begriff der Vereinbarkeit von Hypothesen mit Beobachtungen näher bringen und zum anderen die übliche Deutung von Überdeckungswahrscheinlichkeit als langfristiger Erfolgsrate auf einen qualitativen Sicherheitsindex abschwächen. Die übliche Häufigkeitsdeutung von den genannten Größen ist sowieso nur eingeschränkt zu verstehen als *bedingte* Fehler (und deren Wahrscheinlichkeiten) – bedingt auf das Zutreffen von unterstellten Modellen.

Informelle Inferenz

Welche Schlüsse können wir aus einer kleinen Stichprobe ziehen über den Wert eines unbekannten Anteils q an Meinung A in der Grundgesamtheit? Wir unterstellen, dass wir den Anteil q kennen, und ziehen eine Stichprobe von 10 Personen. Für jede Person, die Meinung A hat, notieren wir eine 1; wenn sie die gegenteilige Meinung nicht-A vertritt, so schreiben wir dafür 0 („weiß nicht“ oder keine Antwort lassen wir außer Acht). Wir ziehen wiederholt Stichproben und untersuchen, was passiert.

Hier ist eine kleine Bemerkung hilfreich: Eigentlich hat man immer nur eine *einzige* Stichprobe (hier vom Umfang 10). Wie gut diese eine Stichprobe ist, wie gut eine Stichprobe einen Sachverhalt einschätzen lässt, validiert man damit, dass man diese eine Stichprobe öfters wiederholt. Dieser Prozess des wiederholten Ziehens von Stichproben ist in der Praxis völlig abwesend. Man hat nur eine. Wenn man mehr Ressourcen (Geld, Zeit) hat, macht man diese eine Stichprobe größer aber man zieht doch nicht wiederholt eine kleine Stichprobe. Das machen wir hier auch nur aus didaktischer Absicht. Und wir können das, weil wir so tun, als vergleichen wir das beobachtete Ergebnis (jetzt wohl besser gesagt, das in Frage kommende Ergebnis – nämlich 70% und mehr mit Meinung A in der Stichprobe zu bekommen) mit fiktiven Werten für die Grundgesamtheit.

Diese Denkweise ist für die Beurteilende Statistik typisch: Immer werden beobachtete Werte mit auf Hypothesen beruhenden Szenarien für die Grundgesamtheit verglichen. In Abb. 11 wiederholen wir eine Stichprobe vom Umfang 10 insgesamt 600 Mal. Daraufhin untersuchen wir mit Mitteln der beschreibenden Statistik diese hypothetische Vielfalt und erkennen, dass bei einem (unterstellten) Wert von q = 0,39 ein Stichprobenanteil von mindestens 70% mit Meinung A immerhin eine Wahrscheinlichkeit von 0,052 hat. Es ist zwar selten aber nicht ganz ungewöhnlich, dass Meinung A trotz geringen Anteils (0,39 oder 39%) in der Population in der Stichprobe mit 70% und mehr vertreten ist.

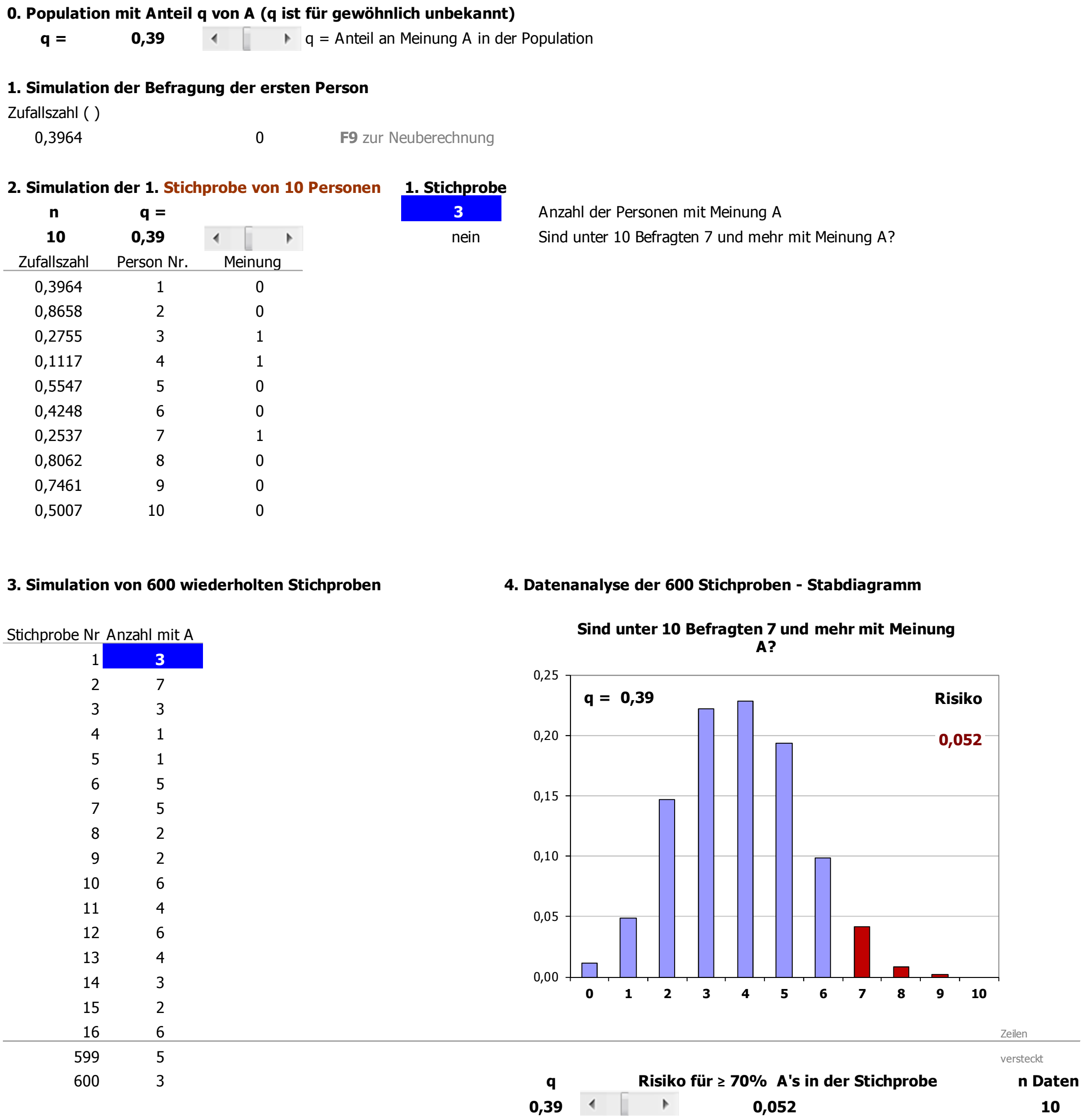

Abb. 11: Die verschiedenen Stufen der Simulation: 0. die Grundgesamtheit, 1. die Auswahl einer Person, 2. die Auswahl einer Stichprobe von 10 Personen, 3. die wiederholte Auswahl einer Stichprobe von 10 Personen.

Wir verändern den Wert von q dynamisch mit einem Regler und erkennen, dass alle Werte größer als 0,39 auch mit einer Beobachtung von 7 oder mehr (zumindest 70%) Meinung A in den Stichproben *kompatibel* sind. Gleichzeitig sehen wir, dass Werte von $q < 0{,}39$ mit einer Beobachtung von 70% und mehr in der Stichprobe nicht mehr kompatibel sind. Diese Simulation ist eine Vorbereitung auf den statistischen Test sowie die Interpretation von p-Werten (p values, normalerweise würden wir Anteile mit eben diesem Buchstaben p bezeichnen; um diesen Konflikt in der Bezeichnung zu vermeiden, weichen wir für den Anteil auf q aus). p-Werte werden neuerdings vielfach anstelle der Fehlerwahrscheinlichkeit 1. Art (α) verwendet. Aufgrund einer Beobachtung 7 (von 10) hat die Hypothese 0,39 für den Anteil an A einen p value von 0,052. Normalerweise betrachten wir eine Hypothese als abgelehnt, wenn ihr p-Wert aufgrund einer Beobachtung kleiner als 5% (bzw. 1%) ist.

Noch wichtiger als die Interpretation von p-Werten ist uns hier jedoch die Einsicht, dass das vollkommen analoge Szenario, jetzt aber mit Stichproben vom Umfang 100 zu einer wesentlich schärferen Information führt: jetzt sind – wie man aus Abb. 12 sehen kann, bereits Anteile von 0,60 für die Meinung A in der Grundgesamtheit nicht mehr kompatibel mit einer Beobachtung von 70% und mehr in der Stichprobe. Mit anderen Worten: Wäre die Meinung in der Grundgesamtheit mit 0,60 vertreten, so ist es sehr unwahrscheinlich (weniger als 0,020 siehe Abb. 12), dass eine Stichprobe ein so extremes Ergebnis wie 70% oder mehr ergibt. Angesichts einer solchen Beobachtung hat die Hypothese $q = 0{,}60$ eine zu kleine stochastische Erklärungskraft.

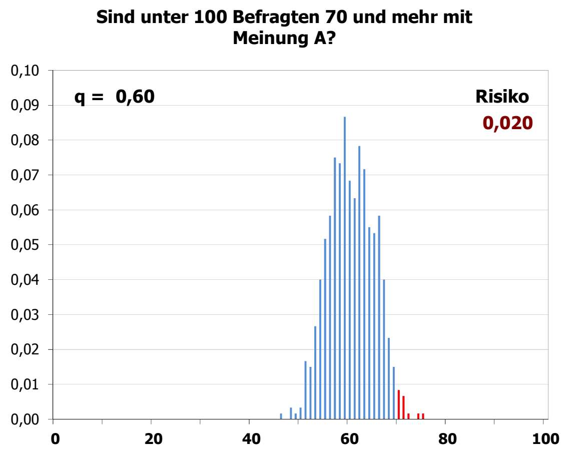

Abb. 12: Risiko, bei 60% A in der Population in einer Stichprobe mit 100 Personen 70% und mehr A zu finden.

Konfidenzintervall

Ein 95% Konfidenzintervall „deckt mit 95%iger Sicherheit den gesuchten Parameter ab“. Ziel dieser Simulation ist es, den Unterschied aufzuzeigen der entsteht, wenn ein Konfidenzintervall aus 5 oder aus 20 Daten berechnet wird. Dazu werden zu 5 und zu 20 Daten jeweils 200 Stichproben simuliert und die Abweichung zum tatsächlichen Wert (hier 0) in einem Diagramm dargestellt.

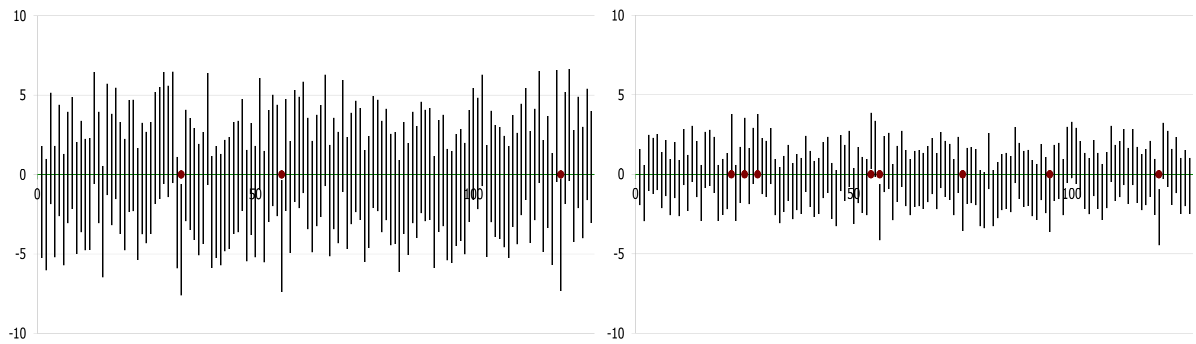

Abb. 13: Ausschnitt aus den wiederholten Konfidenzintervallen – links bei 5, rechts bei 20 Daten: Erhöhung der Präzision bei Vergrößerung der Stichprobe anstelle der Überdeckungswahrscheinlichkeit als langfristige Erfolgsrate.

Die Punkte auf der Achse (in Abb. 13) symbolisieren Intervalle, die den Parameter nicht überdecken und machen die Erfolgsrate augenfällig. Erneuert man die gesamte Simulation mehrmals, sieht man, dass einerseits die Erfolgsrate in beiden Fällen immer um die 95% schwankt, andererseits, und das ist noch wesentlicher, dass die Intervalle basierend auf 20 Daten nur halb so lang sind wie mit 5 Daten. Interpretiert man die Überdeckungswahrscheinlichkeit als qualitativen Index – im Sinne von „je höher, desto besser”, so kann man erkennen, dass eine Vervierfachung des Stichprobenumfangs die Länge und damit die Unschärfe der auf dem Konfidenzintervall basierenden Aussage über den Wert des (normalerweise unbekannten) Parameters halbiert.

6. Zusammenfassung

Mit dem Anspruch, den Modellbildungsgedanken in den Unterricht einzubringen und damit die Vielfalt der Begriffe erst wirklich entstehen zu lassen, wird auch die Mathematik vielfältiger und schon daher auch schwieriger. Man muss ja mehr Modelle parallel durchrechnen und verstehen. Andererseits wird die Mathematik in einen Kontext „realer“ Anwendung gestellt und man kann erkennen, wie die Modellbildung die Situation (die Entscheidung und deren Folgen) verbessert. Die Begriffe können nicht mehr nur schematisch angewendet werden, eine sinnvolle Interpretation muss gefunden werden. Damit man die Mathematik halbwegs unter Kontrolle hält, wird der systematische Einsatz von Software tragend. Anstelle einer Tabellenkalkulation wie EXCEL kann man auch zunehmend die ausgebauten Fähigkeiten von GeoGebra nutzen. (Mit besonders Interessierten könnte man hierzu REXCEL oder gar die Programmiersprache R ausprobieren.)

Mit dem Simulationsansatz kann man nicht nur verschiedenste Modelle durchrechnen, man kann sowohl die Modellbedingungen als auch die Schlüsse vollständig graphisch darstellen, was eine umfassendere Sicht auf die Bedingungen bietet. Darüber hinaus kann man theoretische Konzepte auch durch fiktive, simulierte Daten in konkreter Form anbieten und analysieren. Das kann auch komplizierte Aussagen (über Konvergenz, oder über den Sachverhalt, ob einzelne Hypothesen mit bestimmten Beobachtungen kompatibel sind) erhellen. Wenn man dann noch klar stellen kann, dass eine Häufigkeitsdeutung von Wahrscheinlichkeit und anderen tragenden Konzepten nur den Anspruch einer Szenario-Figur hat, dass die genauen Werte von Wahrscheinlichkeiten mehr als Anhaltspunkte denn als exakte Werte aufzufassen sind (jedenfalls wenn man den Bereich der alten Glücksspiele verlässt), ist man auf dem Weg für ein umfassenderes Verständnis angelangt. Der Weg wird immer anziehender, wenngleich er lange ist.

Die Autoren möchten an dieser Stelle Herrn Hans-Stefan Siller für die umsichtige und kritische Diskussion des Artikels herzlich danken.

Literatur

Borovcnik, M. (o.J.): Materialien zur Stochastik. Online: wwwg.uni-klu.ac.at/stochastik.schule/Boro/index_inhalt.

Borovcnik, M. (1988): Korrelation und Regression – ein inhaltlicher Zugang zu den grundlegenden mathematischen Konzepten. In: Stochastik in der Schule 8(1), 5–28.
Online: http://stochastik-in-der-schule.de/sisonline/struktur/jahrgang08-88/heft1/1988-1_Borovc.pdf.

Borovcnik, M. (2009a): Aufgaben in der Stochastik – Chancen jenseits von Motivation. *Schriftenreihe Didaktik der Mathematik der Österr. Math. Ges.* 42, 20–42.
Online: www.oemg.ac.at/DK/Didaktikhefte/2009%20Band%2042/VortragBorovcnik.pdf.

Borovcnik, M. (2011): Strengthening the Role of Probability within Statistics Curricula. In: Batanero, C.; Burrill, G.; Reading, C.; Rossman, A. (Hrsg.): *Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education: A joint ICMI/IASE Study*. New York: Springer.

Borovcnik, M., Kapadia, R. (2011): Modelling in Probability and Statistics–Key ideas and innovative examples. In: Maaß, J.; O‘Donoghue, J. (Hrsg.): *Real-World Problems for Secondary School Students–Case Studies*. Rotterdam: Sense Publ.

Dubben, H.-H.; Beck-Bornholdt, H.-P. (2007): Der Hund, der Eier legt. 3. veränderte Aufl. Reinbek bei Hamburg: Rowohlt.

Eichler, A.; Vogel, M. (2010). Daten und Zufall als einende Leitidee. *Paper presented at the DAGStat2010*, Dortmund, 25.03.2010.

Kataoka, V.Y. et al (2009): Probability Teaching in Basic Education in Brazil: Assessment and Intervention. Proc. of TSG 13 on „Research and development in the teaching and learning of probability”. Intern. Congress on Mathematical Education 12 (ICME 12). Online: http://tsg.icme11.org/tsg/show/14#prog.

König, G. (2011): Motivation bei Wettbewerben: Stochastische Aspekte einer Diskussion in der Zeitschrift „Psychological Science“. In: *Stochastik in der Schule* 31(3), 22–25.

Lorenz, R.J. (1996): *Biometrie*. Stuttgart: G. Fischer.

Schnell, S, (2011): „Das ist doch kein Zufall” – Entwicklung von Schülervorstellungen zum empirischen Gesetz der großen Zahlen. Vortrag auf dem AK „Stochastik”, Kassel, 24.09.2011.