Die Berechnung von $P(X + Y = w)$ mittels Tabellenkalkulation

John C. Turner, Annapolis/USA
Übersetzung: Ingeborg Strauß, Kronberg im Taunus

Zusammenfassung: Benutzt man die Spreadsheet-Funktion SUMPRODUCT, hinter der sich die Skalarprodukt-Bildung verbirgt, kann man die Wahrscheinlichkeiten der Summen unabhängiger diskreter Zufalls-Variablen berechnen lassen. Dadurch können Schüler spezielle Eigenschaften der Summen von Binomial- und Poisson-Verteilungen bestätigen. Auch liefert sie eine Methode, um die Verteilung der Summe zweier oder mehrerer beliebiger Zufalls-Variablen zu berechnen. Zusätzlich gibt sie dem Schüler ein Hilfsmittel an die Hand zur Berechnung der Wahrscheinlichkeiten aus der Differenz von Zufalls-Variablen und damit der Wahrscheinlichkeit, dass eine Variable (um einen gewissen Betrag) „besser" ist als die andere.

Einleitung

Das Aufsummieren von Zufallsvariablen umgeht man meist in (Einführungs-) Kursen zur Wahrscheinlichkeitsrechnung. Spezialfälle, wie die Tatsache, dass die Summe von Poisson-verteilten Variablen eine Poisson-Verteilung hat, werden ab und zu angesprochen. Der *Zentrale Grenzwertsatz* liefert dann eine Approximation, wenn die Anzahl der Zufallsvariablen genügend groß ist. Andere interessante Probleme werden dagegen ausgespart, wohl weil die Berechnung ziemlich langweilig und langwierig ist. Hier nun will ich eine Methode vorstellen, die durch Einsatz eines Tabellenkalkulations-Programms die Berechnung der exakten Verteilung der Summe zweier (oder mehrerer) diskreter Zufallsvariablen beliebiger Verteilungen gestattet. Sodann werde ich zeigen, wie man diese Methode bei der Untersuchung verschiedener interessanter Aufgabenstellungen auch in Grundkursen einsetzen kann.

Das Problem

Angenommen, X und Y seien unabhängige diskrete Zufallsvariablen mit den resp. Wahrscheinlichkeitsmaß-Funktionen P_x und P_y. Es sei $W = X + Y$ und bezeichne die Wahrscheinlichkeitsmaß-Funktion für W bei P_w. Dann gilt

$$P_W = \sum_{x+y=t} P_x(x) \cdot P_y(y)$$

Selbst Anfänger können diese Gleichung leicht verifizieren. Zunächst betrachte man alle Fälle mit $x+y=t$. Da $P_x(x)$ und $P_y(y)$ die zu x und y gehörenden Wahrscheinlichkeiten angeben und X und Y unabhängig sind, ist das Produkt dieser Wahrscheinlichkeiten gleich der Wahrscheinlichkeit des Paares (x,y). Alle diese Paare sind disjunkt. Demnach ist die Gesamtwahrscheinlichkeit für alle Paare durch die Summe der Wahrscheinlichkeiten für jedes Paar berechenbar.

Stochastik in der Schule **19**(1999), Nr. 1, S. 14-19
Übersetzung aus Teaching Statistics, vol. 20, no. 2 (Summer 1998), S. 46-48

Das Soda-Beispiel

In meinem Stochastik-Kurs benutzte ich folgendes Beispiel zur Einführung in die Methode: Auf einer von uns veranstalteten Party sind Xavier und Yolanda angeheuert, um die Gäste mit Soda zu versorgen. Tabelle 1 zeigt die Wahrscheinlichkeiten dafür, dass Xavier bzw. Yolanda die angegebene Zahl von Soda-Flaschen tragen. Wir nehmen an, dass die beiden voneinander unabhängig agieren.

Zahl der Flaschen	Xavier	Yolanda
0	0,25	0,15
1	0,30	0,45
2	0,35	0,40
3	0,10	0,00

Tab. 1: Wahrscheinlichkeiten für das Soda-Problem

Wie groß ist die Wahrscheinlichkeit, dass die beiden insgesamt zusammen 3 Soda-Flaschen tragen?

Tabelle 2 führt alle möglichen Kombinationen auf, die zur Gesamtzahl von 3 Soda-Flaschen führen:

Xavier	Yolanda
3	0
2	1
1	2
0	3

Tab. 2: Alle Fälle für $X+Y=3$

In Tabelle 3 sind alle obigen Ereignisse durch ihre zugehörigen Wahrscheinlichkeiten ersetzt. Zusätzlich sind die Produkte dieser Wahrscheinlichkeiten berechnet und aufsummiert worden. Also ist die Wahrscheinlichkeit für genau 3 Soda-Flaschen gleich 0,2925.

Xavier	Yolanda	Produkt
0,10	0,15	0,0150
0,35	0,45	0,1575
0,30	0,40	0,1200
0,25	0,00	0,0000
	Summe: 0,2925	

Tab. 3: Berechnung von $P(X+Y=3)$

Für irgendeine andere Anzahl von Flaschen verläuft die Rechnung analog. Der einzige Unterschied ist, dass eine der Spalten in Tabelle 2 oder 3 so aufwärts oder abwärts wandert, dass die Werte in jeder Zeile von Tabelle 3 das ihre zur gewünschten Summe beitragen. Möglicherweise werden also Werte ergänzt werden müssen, die nicht in der originalen Tabelle 2 vorhanden sind. Zum Beispiel: Um die Gesamt-Wahrscheinlichkeit für 1 Flasche zu finden, wie in Tabel-

le 4 demonstriert, müssen wir zu $X = 3$ die Ergänzung $Y = -2$ setzen, denn das summiert sich zu 1. Es macht keinen Sinn, $Y = -2$ zu betrachten (es sei denn, Yolanda räumt leere Flaschen ab), weshalb wir diesem Ereignis die Wahrscheinlichkeit 0 zuordnen. Man beachte, dass Tabelle 1 solch einen unmöglichen Fall für $Y = 3$ enthält.

Xavier		Yolanda		
Anzahl	Wahrscheinlichkeit	Anzahl	Wahrscheinlichkeit	Produkt
3	0,10	- 2	0,00	0,0000
2	0,35	- 1	0,00	0,0000
1	0,30	0	0,15	0,0450
0	0,25	1	0,45	0,1125
	0,00	2	0,40	0,0000
	0,00	3	0,00	0,0000
				Summe: 0,1575

Tab. 4: Berechnung von $P(X+Y=1)$

Zusammengefasst leistet die Methode folgendes:

1. Man ergänze in Tabelle 1 unter und über den gegebenen Wahrscheinlichkeiten die Wahrscheinlichkeitswerte 0.
2. Man invertiere die erste Spalte, so dass die Eintragungen von unten nach oben verlaufen.
3. Man verschiebe die beiden Spalten so, dass die Ereignis-Zahlen-Paare in jeder Zeile sich zu dem gewünschten Wert summieren.
4. Man multipliziere die korrespondierenden Wahrscheinlichkeiten in jeder Zeile und summiere sie.

Lösung mit Tabellenkalkulation

Wenn die Schüler bereits mit der Vektor-Rechnung vertraut sind, sollte man sie darauf hinweisen, dass Schritt 4 nichts anderes ist als das Skalar-Produkt der „Gleit"spalten. Spreadsheets wie beispielsweise Microsoft Excel oder Corel's Quattro Pro besitzen eine eingebaute Funktion, die das Skalar-Produkt berechnet. Sie wird über SUMPRODUCT(block1,block2) aufgerufen.

Tabelle 5 verdeutlicht die Vorgehensweise für das Soda-Problem. Die Spalten A bis D enthalten die individuellen Werte für X und Y sowie die zugehörigen Wahrscheinlichkeiten. (Man beachte, dass die Nullen in den Spalten A und C in derselben Zeile stehen.) Zelle F2 basiert auf der Formel SUMPRODUCT(B$2:B$5,D2:D5). Das Dollar-Zeichen in der ersten Adressierung macht die Angaben absolut, das Fehlen des Dollar-Zeichens in der zweiten Adressierung dagegen relativ. Wird nun F2 nach F3 kopiert, wird die zweite Adressierung angepasst zu D3:D6, etc. Das heißt, F2 enthält das Skalar-Produkt der Paare für $X+Y = 0$, F3 dagegen die Gesamt-Wahrscheinlichkeit für $X+Y=1$. Das Ergebnis

0,2925 in F5 ist von den obigen Erläuterungen her bekannt. Es sei dem Leser überlassen zu verifizieren, dass die Summe all dieser Einträge 1 ergibt. Spalte E listet die Werte von $X+Y$ auf, für die die Wahrscheinlichkeiten in Spalte F errechnet sind. E2 enthält die Formel =A$2+C2, die dann in die daruntersteheden Zellen kopiert wurde.

	A	B	C	D	E	F
1	X	$P(X)$	Y	$P(Y)$	$X+Y$	$P(X+Y)$
2	3	0,10	- 3	0,00	0	0,0375
3	2	0,35	- 2	0,00	1	0,1575
4	1	0,30	- 1	0,00	2	0,2875
5	0	0,25	0	0,15	3	0,2925
6	- 1	0,00	1	0,45	4	0,1850
7	- 2	0,00	2	0,40	5	0,0400
8						
9						
10						
11						

Tab. 5: Tabellenblatt für $P(X+Y)$

Wer anstelle von Excel das Programm Quattro Pro einsetzt, muss beim Aufruf von SUMPRODUCT folgendes berücksichtigen: Ist irgendeine relevante Zelle leer, wird ERR zurückgegeben. Es ist also sicherzustellen, dass alle diese Zellen mit Nullen gefüllt werden.

Tabelle 5 kann die Grundlage sein für ein verallgemeinertes Spreadsheet, das eine ganze Reihe weiterer Probleme zu behandeln in der Lage ist. Wie bei allen Tabellenblättern ist es auch hier so, dass bei Veränderung eines Input-Wertes automatisch die Outputs neu berechnet werden. Für irgendeine andere Menge von Wahrscheinlichkeiten genügt es also, die alten Werte durch die neuen zu überschreiben. Die Gesamtwahrscheinlichkeiten für $X+Y$ werden dann automatisch angepasst. Dies gilt natürlich nur solange, wie die Anzahl neuer Werte die der alten nicht übersteigt. Hat man weniger Werte zur Verfügung, ersetzt man einfach die überzähligen durch 0. Das Ausgangs-Spreadsheet sollte also so umfangreich wie möglich sein.

Anwendungen

Auf der Grundlage eines solchen nicht zu kleinen Tabellenblattes unter Einsatz der in solcher Software implementierten Funktionen für die gängigsten diskreten Wahrscheinlichkeits-Maß-Funktionen ist es ein leichtes, die bekannten Resultate für die Summen von Binomial-Verteilungen mit gleichem p zu berechnen, oder die Summen von Poisson-Verteilungen, oder die Summen von negativen Binomial-Verteilungen mit gleichem p. Weiterhin ist es möglich, die Verteilung der

Summe zweier Binomial-Verteilungen mit unterschiedlichen p's zu betrachten, was zu einem interessanten Resultat führt.

Angenommen, X ist binomial-verteilt mit den Parametern N und p. Weiter angenommen, Y ist ebenfalls binomial-verteilt mit demselben N und der Erfolgs-Wahrscheinlichkeit $1 - p$. Das heißt beispielsweise, wenn die Erfolgs-Wahrscheinlichkeit für X 0,7 beträgt, ist sie für Y gleich 0,3. Es sei $W = X+Y$. Dann ist die durchschnittliche Erfolgswahrscheinlichkeit für W gleich 0,5. Dies ist der zugehörige Wert für p, wenn W binomial-verteilt ist. Werden die Wahrscheinlichkeiten für X und Y in ein Spreadsheet analog Tabelle 5 eingetragen, zeigt sich, dass die Wahrscheinlichkeits-Maß-Funktion für W symmetrisch ist. Die einzige symmetrische Binomial-Verteilung ist die mit $p = 0{,}5$. Vergleicht man die Wahrscheinlichkeiten auf dem Tabellenblatt mit denen der Binomial-Verteilung für $p = 0{,}5$, sieht man dagegen interessante Unterschiede. Erstens, die beiden Mengen von Wahrscheinlichkeiten stimmen nicht überein. Das ist noch nicht besonders bemerkenswert, denn wir wissen, dass Summen von Binomial-Verteilungen nur bei gleichen p's wieder binomial-verteilt sind. Die zweite Beobachtung ist überraschender: Gleich welches p wir dem X zuweisen, es ergibt sich eine W-Verteilung mit ausgeprägterem „Peak" als bei der symmetrischen Binomial-Verteilung.

Der Grund ist leicht einzusehen. Zunächst einmal ist zu konstatieren, dass die Resultate bezüglich p (hier 0,3) und $1 - p$ (hier 0,7) symmetrisch sind, denn es werden einfach nur die Rollen von X und Y vertauscht. Betrachten wir weiterhin die Extrem-Fälle $p = 0$ (oder $p = 1$), dann kann X nur 0 und Y nur N sein, womit W auch nur gleich N sein kann. Daraus ergibt sich natürlich eine stark Gipfel–geprägte Verteilung. Es ist unwahrscheinlich, dass diese Eigenart ohne den Einsatz eines Tabellenblattes bemerkt worden wäre.

Differenz von Zufalls-Variablen

Ein weiteres Problem, das im allgemeinen im Schul-Alltag umgangen wird, ist das der Verteilung der Differenz zweier binomial-verteilter Zufallsvariablen mit unterschiedlichen p's. Auch hierbei ist ein Spreadsheet hilfreich und leicht einsetzbar. Als motivierendes Beispiel betrachten wir einen Multiple-Joice-Test zweier Bewerber. Der eine von beiden möge mehr richtige Antworten gegeben haben als der andere. Angenommen, Kandidat A habe eine Wahrscheinlichkeit von 0,6 für das richtige Erraten der Antworten, Kandidat B nur eine solche von 0,5. Wie groß ist die Wahrscheinlichkeit, dass dennoch B seinen Kontrahenten ausstechen und damit den Job bekommen wird?

Die Frage ist äquivalent der für $P(B - A > 0)$. Wieder können wir unser obiges Tabellenblatt verwenden, nur dass jetzt die *Differenzen* in jeder Zeile den gewünschten Wert haben müssen. Die Wahrscheinlichkeiten für B werden also in der bekannten Weise eingetragen, sie nehmen den Platz unter Y ein. Für Kandidat A verfahren wir entsprechend. Die Wahrscheinlichkeiten werden analog B

abwärts gehend aufgeführt. Diese Anordnung kann man auf zweierlei Art interpretieren. Verschieben wir die Spalten auf- oder abwärts, behalten die Werte für A und B eine konstante Differenz. Man kann aber auch die Differenzenbildung als Summe aus $B + (-A)$ auffassen. Dann schreiben wir die Wahrscheinlichkeiten abwärts gelesen auf und assoziieren sie mit den entsprechenden Werten von $-A$. Letztere Betrachtungsweise ist in Tabelle 6 realisiert:

	A	B	C	D	E	F	G
1							
2							
3							
4							
5						0	
6	3				- 3	0,027	
7	2				- 2	0,135	
8	1				- 1	0,279	0,441
9	0	0,064	0	0,125	0	0,305	0,305
10	- 1	0,288	1	0,375	1	0,186	0,254
11	- 2	0,432	2	0,375	2	0,060	
12	- 3	0,216	3	0,125	3	0,008	
13						0,000	

Tab. 6: $P(B-A)$

Tabelle 6 illustriert die Berechnungen für den Fall $N = 3$. Die Wahrscheinlichkeiten in Spalte B sind binomial-verteilt mit $p = 0{,}6$, die in Spalte D mit $p = 0{,}5$. Man beachte die Reihenfolge in Spalte E. Die Differenz zwischen Kandidat A und Kandidat B variiert zwischen – 3 und + 3. Die Werte in Spalte G repräsentieren die Wahrscheinlichkeit, dass $B - A$ negativ ist (0,441 = 0,027 + 0,135 + 0,279), dass $B = A$ ist (0,305) und dass $B - A$ positiv ist (0,254). Für $N = 3$ ergibt sich demnach eine 25%-Chance dafür, dass der schwächere Bewerber „siegt".

Um die Werte in Tabelle 6 anzupassen, war es notwendig, das Tabellenblatt zu modifizieren. Die Formel in Zelle F9 lautet nun SUMPRODUCT(B$6:B$12, D6:D12). Die Adressierungsbereiche wurden ausgedehnt, um die zusätzlichen Wahrscheinlichkeiten zu erfassen. Zusätzlich mussten mehrere leere Zeilen vorgeschaltet werden. Damit ist sichergestellt, dass die Formel in F5 nicht auf Zellen ausserhalb des Tabellenblattes zugreift.

John C Turner
US Naval Academy, Annapolis, USA.
e-Mail: jct@usna.navy.mil

Anm. zu S. 17 unten: Eine diskrete Zufallsgröße X heißt *negativ binomialverteilt* mit den Parametern p und v ($0 < p < 1$, $v > 0$), wenn für die Einzelwahrscheinlichkeiten gilt: $P(X=m)=\binom{-v}{m}(-p)^m(1-p)^v$, $(m=0,1,2,...)$

Für ganzzahlige v gibt X die Anzahl der Misserfolge vor dem v-ten Erfolg beim Bernoulli-Schema an. *I. St.*