Stochastik in der Schule SiS

Zeitschrift des Vereins zur Förderung des schulischen Stochastikunterrichts

Inhaltsverzeichnis Heft 2, Band 35 (2015)

Vorwort des Herausgebers

Liebe Leserin, lieber Leser,

haben Sie schon einmal die Freude gehabt, in einen Senfkrapfen zu beißen? Haben Sie sich schon einmal gefragt, wie man den Textanteil in einer Zeitung schätzen kann? Stochastik lebt von interessanten Problemstellungen und den anschließenden Modellierungen zur Lösungsfindung. Hierzu finden Sie in diesem Heft konkrete Anregungen für Ihren Unterricht.

Über den *p*-Wert ist schon viel geschrieben worden. Lassen Sie sich überraschen. Sie werden hier bestimmt etwas Neues finden.

Bootstrap-Konfidenzintervalle – sind die etwas für die Schule? Auch hierzu finden Sie in diesem Heft eine eindeutige Beantwortung.

Ein übersetzter Artikel aus „Teaching Statistics“ beschreibt die Vorteile einer Einführung des Variationskoeffizienten in der Beschreibenden Statistik.

Wenn Sie noch weitere Ideen benötigen, müssen Sie nur die Rezension zu einem gehaltvollen Buch lesen.

Dank an G. König für seine Bibliographische Rundschau. Besonders möchte ich auch auf die beiden Einladungen hinweisen.

Ich wünsche den Leserinnen und Lesern beim Lesen des Heftes viel Freude und viele Anregungen.

Hannover, April 2015

Reimund Vehling

Senfkrapfen

FRIEDRICH BARTH & RUDOLF HALLER, MÜNCHEN

Zusammenfassung: *Von N gleich aussehenden Objekten seien S Objekte defekt. Ein Prüfer entnimmt ein Objekt auf gut Glück um festzustellen, ob es defekt ist. Der Hersteller kann die Objekte dem Prüfer auf verschiedene Art präsentieren. Hängt die Wahrscheinlichkeit dafür, dass der Prüfer ein defektes Stück entdeckt, von der Art der Präsentation ab? Diese Situation wird im Beitrag beispielhaft und anschaulich vorgeführt. Sie lässt sich auch im Unterricht nachspielen, indem man geeignete gleich aussehende Objekte verwendet, von denen einige verändert sind, ohne dass dies unmittelbar erkennbar ist.*

Eine Faschingseinladung mit Überraschung

Zita[1] lädt zu einem Faschingsabend[2] mehrere Freunde ein, unter ihnen auch Xaver[3], dem sie einen Streich spielen will. Dazu bäckt sie N Krapfen[4], von denen sie S Stück mit Senf statt mit Marmelade füllt ($1 \le S < N$). Da sie in der Schule Stochastik gelernt hat, überlegt sie, auf welche Art und Weise sie ihre Krapfen auf verschiedenen Tellern arrangieren kann, damit Xaver, der als Erster einen Krapfen nehmen darf, höchstwahrscheinlich in einen Senfkrapfen beißt. Xaver wählt auf gut Glück einen Teller und nimmt sich von diesem Teller einen Krapfen.

1. Art. Zita legt alle Krapfen auf einen Teller. Mit welcher Wahrscheinlichkeit $P(X)$ tritt das Ereignis $X :=$ »Xaver greift sich einen Senfkrapfen« ein? Welcher Wert ergibt sich für $P(X)$ für $N = 18$ und $S = 3$?

Lösung. $P(X) = \frac{S}{N}$.

Speziell $P(X) = \frac{3}{18} \approx 16{,}7\ \%$.

2. Art. Zita arrangiert ihre 18 Krapfen auf drei Tellern zu je sechs so, dass auf dem ersten Teller ein Senfkrapfen, auf dem zweiten zwei und auf dem dritten kein Senfkrapfen liegt.

Abb.1: Die Verteilung von 3 Senfkrapfen und 15 Krapfen auf drei Teller gemäß der **2. Art**

Mit welcher Wahrscheinlichkeit greift Xaver sich einen Senfkrapfen?

$$P(X) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{2}{6} + \frac{1}{3} \cdot 0 = \frac{1}{6} \approx 16{,}7\ \%.$$

Verallgemeinerung

2V. Zita legt auf n Teller jeweils gleich viele Krapfen. Dazu muss die Anzahl N der gebackenen Krapfen ein Vielfaches von n sein. Zeige: Die Wahrscheinlichkeit, dass Xaver einen Senfkrapfen nimmt, ist unabhängig von der Verteilung der S Senfkrapfen auf die n Teller.

Lösung. Auf den i-ten Teller werden $\frac{N}{n}$ Krapfen, darunter s_i Senfkrapfen gelegt, wobei $S = \sum_{i=1}^{n} s_i$ ist.

Dann gilt $P(X) = \frac{1}{n} \sum_{i=1}^{n} \frac{s_i}{\frac{N}{n}} = \frac{1}{N} \sum_{i=1}^{n} s_i = \frac{S}{N}$,

d. h., $P(X)$ hängt nicht von der Verteilung der Senfkrapfen auf die Teller ab.

3. Art. Zita hat sieben Krapfen gebacken und füllt vier davon mit Senf. Sie verteilt sie so auf zwei Teller, dass auf jedem Teller mindestens ein Krapfen liegt.

a) Welche Arrangements kann Zita vornehmen?

b) Mit welcher Wahrscheinlichkeit beißt Xaver jeweils in einen Senfkrapfen? Bei welcher Anordnung wird diese Wahrscheinlichkeit maximal?

Lösung

a) Ohne Beschränkung der Allgemeinheit legt Zita, da die Teller ja vertauschbar sind, auf Teller A höchstens drei Krapfen und damit auch höchstens drei Senfkrapfen. Damit ergeben sich die folgenden Arrangements (siehe Tabelle 1).

b) Liegen auf Teller A k Krapfen ($1 \le k \le 3$), wovon s Krapfen Senfkrapfen sind ($0 \le s \le k$), dann erhält man aus dem Baumdiagramm (Abbildung 2)

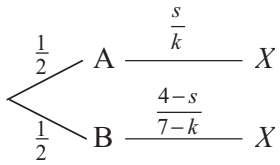


Abb. 2: Baum zur Berechnung der Wahrscheinlichkeit $P(X)$ für den Fall der Präsentation **3. Art**

$$P(X) = \frac{1}{2} \cdot \frac{s}{k} + \frac{1}{2} \cdot \frac{4-s}{7-k}.$$

Damit errechnen sich die $P(X)$-Werte von Tabelle 1. Man entnimmt ihr, dass $\max\{P(X)\} = 0{,}75$ ist, falls man einen Senfkrapfen auf Teller A und die drei restlichen Senfkrapfen auf Teller B legt.

Teller A		Teller B		
Anzahl s der Senfkrapfen	Anzahl k aller Krapfen	Anzahl $S-s$ der Senfkrapfen	Anzahl $N-k$ aller Krapfen	$P(X)$
0	1	4	6	$\frac{1}{3} \approx 0{,}33333$
0	2	4	5	$\frac{2}{5} = 0{,}40000$
0	3	4	4	$\frac{1}{2} = 0{,}50000$
1	1	3	6	$\frac{3}{4} = 0{,}75000$ **Maximum**
1	2	3	5	$\frac{11}{20} = 0{,}55000$
1	3	3	4	$\frac{13}{24} \approx 0{,}54167$
2	2	2	5	$\frac{7}{10} = 0{,}70000$
2	3	2	4	$\frac{7}{12} \approx 0{,}58333$
3	3	1	4	$\frac{5}{8} = 0{,}62500$

Tab. 1: Mögliche Arrangements für die **3. Art** samt $P(X)$

Für die Wahrscheinlichkeit $P(X)$ ergibt sich der dreidimsenionale Graph aus Abbildung 3.

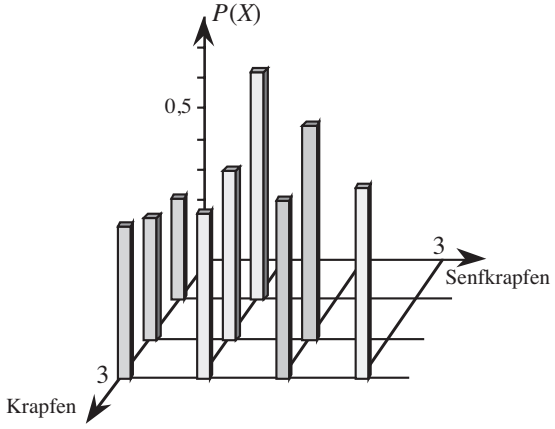


Abb. 3: Die Wahrscheinlichkeit $P(X)$ in Abhängigkeit von der Belegung von Teller A mit k Krapfen, darunter s Senfkrapfen, $1 \le k \le 3 \wedge 0 \le s \le k$

Verallgemeinerung

3V. Zita verteilt ihre N Krapfen, darunter S Senfkrapfen, so auf zwei Teller, dass auf jedem Teller mindestens ein Krapfen liegt. Zeige: Die Wahrscheinlichkeit, dass Xaver in einen Senfkrapfen beißt, wird maximal, wenn Zita auf Teller A als einzigen Krapfen einen Senfkrapfen und auf Teller B die restlichen Krapfen legt.

Lösung. Offensichtlich muss $N \ge 2$ sein. Auf Teller A liegen s Senfkrapfen und insgesamt k Krapfen; auf Teller B liegen dann $S-s$ Senfkrapfen und $N-k$ Krapfen. Ohne Beschränkung der Allgemeinheit sei $k \le N-k$, da ja die Teller vertauscht werden können. Dann gilt

$$0 \le s \le S,\ 1 \le k \le \frac{1}{2}N,\ s \le k.$$

Das Baumdiagramm aus Abbildung 4 liefert

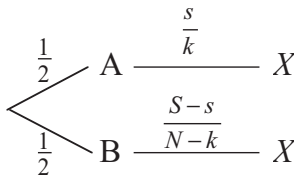


Abb. 4: Baum zur Berechnung der Wahrscheinlichkeit $P(X)$ für den Fall **3V**

$$P(X) = \frac{1}{2} \cdot \frac{s}{k} + \frac{1}{2} \cdot \frac{S-s}{N-k}.$$

$P(X)$ ist eine Funktion der zwei ganzzahligen Variablen s und k, die aus endlichen Intervallen stammen. Maximal gibt es $\frac{1}{2}N(S+1)$ Punkte $(s|k)$, auf denen $P(X)$ definiert ist. Wir müssten also schreiben $P(X) = P_{s,k}(X)$.

Wir bestimmen nun das Maximum von $P(X)$ nach einem Vorgehen, das wir anschaulich wie folgt beschreiben können:

Sucht man den größten Bewohner eines Wohnhauses, dann bestimmt man zuerst den größten Bewohner in jedem einzelnen Zimmer. Aus den Zimmer-Größten ermittelt man anschließend den Haus-Größten.

Auf unser Problem übertragen: Wir halten zunächst k fest und bestimmen das Maximum von $P_{s,k}(X)$ für jedes konstante k. Im Diagramm von Abbildung 1 betrachtet man also zunächst die beiden Balken für $k=1$, dann die drei Balken für $k=2$ und schließlich die vier Balken für $k=3$ und bestimmt jeweils den relativ größten Balken. Aus diesen drei relativ größten Balken gewinnt man dann den absolut größten Balken.

Zur rechnerischen Behandlung formen wir

$P_{s,k}(X) = \frac{1}{2} \cdot \frac{s}{k} + \frac{1}{2} \cdot \frac{S-s}{N-k}$ um zu

$P_{s,k}(X) = \frac{N-2k}{2k(N-k)} \cdot s + \frac{S}{2(N-k)}$.

Wegen $k \le \frac{1}{2}N$ ist $N - 2k \ge 0$.

1. Fall: $N - 2k = 0$, d. h., $k = \frac{1}{2}N$. Das bedeutet: Auf beiden Tellern liegen gleich viele Krapfen. Dann ist $P(X) = \frac{S}{N}$, wie oben bei **2V** gezeigt.

2. Fall: Für festes k ist der Graph von $P_{s,k}(X)$ in einem $(s|P_{s,k}(X))$-Koordinatensystem eine Punktmenge, die auf einer steigenden Gerade mit der Steigung $m = \frac{N-2k}{2k(N-k)}$ und dem Achsenabschnitt $t = \frac{S}{2(N-k)}$ liegt. Das Maximum der Funktion $P_{s,k}(X)$, die nur für $0 \le s \le k$ definiert ist, ist ein Randmaximum an der Stelle $s = k$. Das jeweilige Randmaximum hat den Wert

$P_{k,k}(X) = 1 - \frac{N-S}{2(N-k)}$.

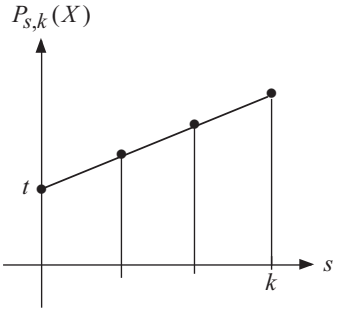


Abb. 5: Graph der Geraden $P_{s,k}(X) = ms + t$

Dieser Ausdruck wird maximal, wenn der zu subtrahierende Bruch minimal wird. Das ist der Fall, wenn der Nenner $N - k$ maximal wird, also k minimal. Dies tritt ein für $k = 1$. Somit gilt:

$\max\{P(X)\} = P_{1,1}(X) = 1 - \frac{N-S}{2(N-1)}$,

d. h.: Auf einen Teller legt Zita als einzigen Krapfen einen Senfkrapfen, alle anderen Krapfen kommen auf den zweiten Teller.

Es ist noch zu zeigen, dass $P_{1,1}(X)$ nicht kleiner als der Maximalwert aus Fall **1** ist.

Annahme:

$\frac{S}{N} > 1 - \frac{N-S}{2(N-1)}, \frac{N-S}{2(N-1)} > \frac{N-S}{N}, N < 2$,

was $N \ge 2$ widerspricht. Die Annahme ist also widerlegt.

Ergebnis: Die Verteilung »Auf einen Teller legt Zita als einzigen Krapfen einen Senfkrapfen, alle anderen Krapfen kommen auf den anderen Teller« liefert maximale Wahrscheinlichkeit für das Ereignis X = »Xaver beißt in einen Senfkrapfen«.

4. Art. Zita verteilt N Krapfen, darunter S Senfkrapfen, so auf die drei Teller A, B und C, dass sie auf jeden Teller mindestens einen Krapfen legt. Falls $S = 1$ ist, legt sie auf Teller A als einzigen Krapfen diesen Senfkrapfen. Falls $S \ge 2$ ist, legt sie auf Teller A und Teller B jeweils einen Senfkrapfen als einzigen Krapfen, den Rest auf Teller C. Zeige: Die Wahrscheinlichkeit, dass Xaver bei diesen Anordnungen in einen Senfkrapfen beißt, ist maximal. – Welchen Wert hat diese maximale Wahrscheinlichkeit bei drei Senfkrapfen unter 18 Krapfen?

1. Fall: $S = 1$

Halten wir die Belegung von Teller C fest, dann wird $P(X)$ nach den Überlegungen bei **3V** maximal, wenn auf Teller A der Senfkrapfen als einziger Krapfen liegt. Es gilt also bei jeder senfkrapfenlosen Belegung von Teller C:

$\max\{P(X)\} = \frac{1}{3}$.

Ergebnis: Man muß den einzigen Senfkrapfen als einzigen Krapfen auf einen Teller legen; die anderen Teller können beliebig mit den Marmelade-Krapfen belegt werden.

2. Fall: $S \ge 2$

Es liegen nicht alle Senfkrapfen auf Teller C. Dann wird $P(X)$ für jede Belegung des Tellers C nach den Überlegungen bei **3V** maximal, wenn auf Teller A ein Senfkrapfen als einziger Krapfen gelegt wird. Wir bezeichnen mit $s|k$ die Belegung eines Tellers mit k Krapfen, von denen s Stück Senfkrapfen sind. Wir halten nun die 1|1-Belegung von Teller A fest und betrachten die Teller B und C. Weil mindestens ein Senfkrapfen noch übrig ist, erhält man das Maximum von $P(X)$ nach den Überlegungen bei **3V** wieder dann, wenn einer dieser Teller, der ohne Beschränkung der Allgemeinheit der Teller B sei, auch eine 1|1-Belegung aufweist. Auf Teller C liegen dann $S - 2$ Senfkrapfen und insgesamt $N - 2$ Krapfen. Damit erhält man mit Hilfe des Baumdiagramms von Abbildung 6

$\max\{P(X)\} = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} \cdot \frac{S-2}{N-2}$,

was sich umformen lässt zu $\frac{2}{3} + \frac{S-2}{3(N-2)}$.

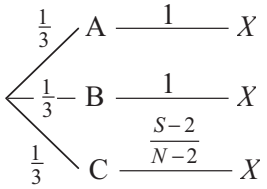


Abb. 6: Baum zur Berechnung der Wahrscheinlichkeit $P(X)$ für den 2. Fall der Präsentation **4. Art**

Ergebnis: Die Verteilung »Auf zwei Teller legt man als einzigen Krapfen einen Senfkrapfen, alle anderen Krapfen kommen auf den dritten Teller« liefert maximale Wahrscheinlichkeit für das Ereignis »Xaver beißt in einen Senfkrapfen«.

Für $N = 18$ und $S = 3$ erhält man

$$\max\{P(X)\} = \frac{2}{3} + \frac{3-2}{3(18-2)} = \frac{11}{16} \approx 68{,}75\,\%\,.$$

5. Art. Zita erinnert sich noch an die optimale Verteilung nach **3V**, nämlich auf Teller A als einzigen Krapfen einen Senfkrapfen zu legen, kann aber ihre Senfkrapfen nicht mehr unter ihren Krapfen identifizieren. Trotzdem legt sie auf Teller A auf gut Glück einen Krapfen und alle anderen auf Teller B. Mit welcher Wahrscheinlichkeit beißt Xaver jetzt in einen Senfkrapfen?

Lösung. Wir zeichnen wieder einen Baum; darin bedeute SA := »Auf Teller A liegt genau ein Senfkrapfen«.

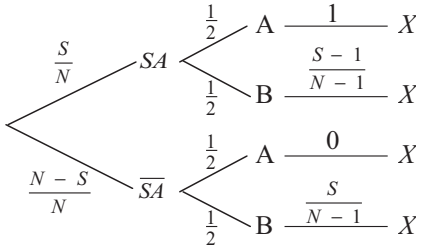


Abb. 7: Baum zur Berechnung der Wahrscheinlichkeit $P(X)$ in Abhängigkeit davon, ob auf Teller A als einziger Krapfen ein Senfkrapfen oder ein Marmeladekrapfen liegt

Aus ihm erhält man

$$\begin{aligned} P(X) &= \frac{S}{N}\cdot\frac{1}{2}\cdot 1 + \frac{S}{N}\cdot\frac{1}{2}\cdot\frac{S-1}{N-1} \\ &\quad + \frac{N-S}{N}\cdot\frac{1}{2}\cdot 0 + \frac{N-S}{N}\cdot\frac{1}{2}\cdot\frac{S}{N-1} \\ &= \frac{S}{2N(N-1)}(N-1+S-1+N-S) \\ &= \frac{S}{2N(N-1)}(2N-2) = \frac{S}{N}. \end{aligned}$$

Das Ergebnis sollte nicht überraschen. Denn jetzt ist ja alles reiner Zufall, wir sind also bei der **1. Art** gelandet.

Epilog. Nach einem derartigen Faschingsabend 2006 scannte Dr. DOMINIK MORHARD vom Institut für klinische Radiologie der Universitätsklinik München verschieden gefüllte Krapfen in einem Computer- und einem Magnetresonanz-Tomographen. Aufgrund der unterschiedlichen Anteile von Wasser, Fett, Öl, Proteinen, Gelier- und Bindemittel konnte er Pudding-, Marmelade-, scharfe und süße Senfkrapfen unterscheiden. So entstand eine wissenschaftliche Publikation.

Für Xaver eignet sich dieses Verfahren nicht; denn es ist aufwändig und kostet mehrere tausend Euro.

Anmerkungen

1 umbrisch = *junges Mädchen*. Die toskanische Dienstmagd ZITA (1210/20–1272) wurde 1696 heiliggeprochen, Fest 27. April.

2 Fasching heißt in manchen Gegenden Deutschlands Karneval oder Fastnacht.

3 baskisch = *neues Haus*. Als Vorname verselbständigt aus dem Namen *Franz Xaver*, der zurückgeht auf FRANCISCO DE JASU Y XAVIER (1506–1552), einen Mitbegründer des Jesuitenordens, päpstlichen Legaten in Indien, 1549 in Japan, 1622 heiliggesprochen; Fest seit 1663 ist der 3. Dezember.

4 Im 9. Jh. bezeichnete das althochdeutsche *krapho* ein hakenförmiges Gebäck, das im Mittelalter *krapfe* hieß. In Berlin und vielen Teilen Ostdeutschlands heißt der bayerische Krapfen *Pfannkuchen*, in vielen Gebieten Nord-, West- und Südwestdeutschlands hingegen *Berliner*.

Literatur

Morhard, D., u. a. (2008): Die diagnostische Wertigkeit von Dual-Energy-CT und 3 Tesla-MRT in der Diagnose von Faschingskrapfen (Berliner Pfannekuchen) – Wo ist die Marmelade, wo der Senf und wo der Pudding? In: *Fortschritte auf dem Gebiet der Röntgenstrahlen und bildgebenden Verfahren,* Heft 4.

Anschrift der Verfasser

Friedrich Barth
Abbachstr. 23
80992 München
e.f.barth@t-online.de

Rudolf Haller
Nederlinger Str. 32a
80638 München
rudolf.haller@arcor.de

„Mit gutem Beispiel vorangehen“ – zum Prinzip des beispielgebundenen Zugangs zur Leitidee Daten und Zufall

MARKUS VOGEL, HEIDELBERG & ANDREAS EICHLER, KASSEL

Zusammenfassung: *Beispiele sind im Mathematikunterricht allgegenwärtig, über sie werden allgemeinere mathematische Einsichten bis in den Einzelfall hinein konkretisiert. In einem datenbasierten Stochastikunterricht ist die beispielhafte Konkretisierung schon durch die Daten und ihren kontextuellen Hintergrund genuin gegeben. In diesem Beitrag werden anhand eines Unterrichtsvorschlags im Themenbereich Konfidenzintervalle didaktische Überlegungen zur Wertigkeit einführender Beispiele angestellt, Kriterien zur didaktischen Beurteilung ihrer unterrichtlichen Gangbarkeit abgeleitet und reflektiert, welche Zugänge zur Leitidee Daten und Zufall durch geeignete Beispiele eröffnet werden können.*

1 Vorweg

„Mit gutem Beispiel vorangehen“ – das ist ein Motto, das sich sprichwörtlich in der Weisheit des Volksmundes niedergeschlagen hat. Mit dem, was damit gemeint ist, lässt sich unserer Auffassung nach aber auch ein didaktisches Prinzip umschreiben, mit dem sich geeignete Einstiege in Themen und Leitideen des Mathematikunterrichts charakterisieren lassen. Daher greifen wir in diesem Beitrag für ein Heft von Stochastik in der Schule darauf zurück, bei dem die Reflexion der zurückliegenden 10 Jahre der Leitidee Daten und Zufall im Vordergrund steht.

Um dem Anspruch dieses Mottos auch in der Struktur dieses Beitrages gerecht zu werden, beginnen wir nach dieser kurzen Einführung direkt in Kapitel 2 mit einer Beispielidee zur unterrichtlichen Umsetzung des Themas Konfidenzintervall. Im anschließenden Kapitel 3 stellen wir anhand des vorausgehenden Beispiels drei Kriterien dar, die bei der Reflexion von geeigneten Beispielen in der Unterrichtsplanung und -auswertung hilfreich sein können. In Kapitel 4 erweitern wir den Fokus, in dem wir über das Leitbeispiel dieses Beitrags hinausgehen, und darlegen, welche Beispiele sich eignen, um im Rahmen eines datenorientierten Stochastikunterrichts Zugänge zu unterschiedlichen Arten von Datenanalysen (vgl. Eichler, 2009; Eichler & Vogel, 2013) zu schaffen. Die vorausgehenden Überlegungen zur didaktischen Bedeutsamkeit von Beispielen werden abschließend in den Zusammenhang der bildungstheoretischen Didaktik eingeordnet.

2 „Warum sind die Fotos so groß?“

Neben verschiedenen anderen Einsichten in Fragen der Gestaltung und Produktion einer Tageszeitung lässt sich mit Blick auf Abbildung 1 unmittelbar die Frage ableiten, wie hoch eigentlich der Textanteil in einer Tageszeitung (oder einer Illustrierten) ist.

In einer offenen Aufgabe formuliert lässt sich daraus ein handlungsorientierter Unterrichtsgang rund um die Themen Stichprobenziehung und Konfidenzintervalle ableiten:

> Bestimmen Sie über eine Stichprobenziehung den Anteil, den der Text in der gesamten Zeitung einnimmt.
>
> Entwickeln Sie entsprechend eine Methode, mit der Sie den Textanteil in einer ganzen Zeitung schätzen. Vergleichen Sie mit dieser Schätzmethode verschiedene Zeitschriften oder Zeitungen.

Im Dialog

Warum sind die Fotos so groß?

Sehr geehrte Damen und Herren, in letzter Zeit kommt es mir vor, als ob ich eine „Bilder-Zeitung“ in der Hand habe und nicht die seit langer Zeit abonnierte Tageszeitung StN. Glauben Sie wirklich, dass Ihre Leser dermaßen großformatige Aufnahmen sehen möchten? Einen interessanten Bericht zu lesen wäre besser, meine ich und bin damit sicher nicht alleine. Oder fehlt es da an Themen?
Frage von [redacted] Stuttgart

Völlig richtig: An interessanten Themen fehlt es nie. Nur würden die ohne Bebilderung kaum Leser finden. Lauter Berichte ohne optische Anreize – das wäre eine unattraktive „Bleiwüste“. Lesen ist schließlich Arbeit, und diese Arbeit sollte so verlockend wie möglich erscheinen, zum Beispiel durch gute Bilder. Schließlich ist das Erste, was der Leser auf einer Seite wahrnimmt, das Foto. Die Größe des Bildes signalisiert: Dieses Thema hat Gewicht. Was jeweils wichtig ist, da können die Meinungen sicher mal auseinandergehen.
[redacted] Ressortleiterin Gestaltung

Abb. 1: Leserbrief aus den Stuttgarter Nachrichten vom 8.3.2014

Hinter dem offenen Aufgabenformat (vgl. z. B. Herget, 2000; Büchter & Leuders, 2005, S. 88 ff.) steht die Vorstellung, dass die Schülerinnen und Schüler Gelegenheit erhalten, selbständig und reflektiert mathematische Eigeninitiative zu entwickeln. Dieses Aufgabenformat, das insbesondere im Gefolge einer Studie der Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (BLK) zur Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts (BLK, 1997) in einem Modul der Weiterentwicklung der Aufgabenkultur im Fach Mathematik erhöhte Aufmerksamkeit gefunden hat, setzt freilich Vorerfahrungen auf Seiten von Lehrkraft und Schülerschaft voraus. Ist dies (noch) nicht gegeben, können differenzierte Arbeitspräzisierungen kurzfristig bei der konkreten unterrichtlichen Umsetzung und langfristig bei der sukzessiven Etablierung einer entsprechenden Unterrichtskultur helfen:

Arbeitsblock 1 (Stichprobenziehung) Bilden Sie zur Bearbeitung Gruppen mit 3–4 Personen:

- Schätzen Sie vor dem Öffnen den Textanteil in der gesamten Zeitung. Halten Sie Ihre Überlegungen und Ihre resultierende Schätzung fest.
- Ziehen Sie mittels eines Lochers eine zufällige Stichprobe von vorab vereinbartem Umfang und ermitteln Sie den Textanteil.
- Schätzen Sie, wie die Textanteile in den übrigen Gruppen und zusammengefasst für die gesamte Gruppe aussehen werden. Begründen Sie!
- Dokumentieren Sie alle in der Gesamtgruppe ermittelten Textanteile in Tabellen- und Diagrammform. Welche Schlussfolgerungen ziehen Sie?

In diesem ersten Arbeitsblock können die Schülerinnen und Schüler mit dem Locher die entscheidende Frage einer Stichprobenziehung erarbeiten: Wie ist zu gewährleisten, dass die Stichprobe wirklich zufällig und damit repräsentativ für die Zeitung wird? Würden beispielsweise nur die Zeitungsränder gelocht, wird sicherlich eine systematische Datenverzerrung zu erwarten sein, da die Ränder nicht bedruckt sind.

In unserem Unterrichtsgang haben wir die Frage so gelöst gesehen, dass die einzelnen Doppelseiten der untersuchten Zeitung zunächst zufällig durchmischt wurden. Anschließend wurde blind eine Doppelseite gezogen, die anschließend ebenfalls zufällig, d. h. nicht entlang der Zeitungskanten und auch nicht parallel bzw. orthogonal dazu, dreimal gefaltet wurde, um anschließend gelocht zu werden. Nach dem Auffalten wurden die ausgestanzten Löcher (auf beiden Seiten) darauf hin ausgezählt, ob durch das Loch ein Buchstabe oder Buchstabenanteil ausgetanzt wurde (bzgl. Überschriften, Werbung oder sonstigen inhaltlichen Kriterien wurde nicht unterschieden) oder nicht (Abb. 2). Anschließend wurde die Seite in den Stapel zurückgelegt und die gesamte Prozedur viermal wiederholt. Das Ergebnis wurde als eine Stichprobenziehung gewertet.

Abb. 2: Beispiel für eine Stichprobenziehung zur Bestimmung des Textanteils in einer Zeitung

Bei acht auf diese Art zustande gekommenen Stichproben ergaben sich dabei folgende Daten für eine Ausgabe der Stuttgarter Zeitung vom 15.02.2014 (Angaben: Stichprobennummer: Anzahl der Löcher – Anzahl Textlöcher – rel. Häufigkeit):

StP_01: 92-39-0,424	StP_02: 84-60-0,714
StP_03: 92-60-0,652	StP_04: 76-41-0,539
StP_05: 88-55-0,625	StP_06: 76-24-0,316
StP_07: 98-30-0,306	StP_08: 96-40-0,417

Bei Aggregierung der Daten ergibt sich bei einer Anzahl von insgesamt 702 Löchern und einer Anzahl von Löchern mit Textanteil von 349 eine relative Häufigkeit von 0,497. Der aggregierten Stichprobe werden die Schülerinnen und Schüler aufgrund des größeren Umfangs (unter Voraussetzung gleicher Ziehungsmodalitäten) größeres Vertrauen schenken. Allerdings stellt sich unserer Erfahrung nach angesichts der Streuung, die sich in der Zusammenschau der einzelnen Stichproben zeigt, auch den Schülerinnen und Schülern unmittelbar die Frage nach der Verlässlichkeit dieser Zahlen und der darauf basierenden Punktschätzung. Damit schließt sich unmittelbar der nächste Arbeitsschritt, die Entwicklung einer Schätzmethode für die Intervallangabe eines Vertrauensbereichs, schlüssig an.

Arbeitsblock 2 (Schätzmethode entwickeln):

In diesem Arbeitsblock geht es darum, eine Methode zu entwickeln, mit der ein Intervall geschätzt werden kann, welches in durchschnittlich 95 % von 100 vergleichbaren Fällen einen zuvor ermittelten relativen Textanteil enthält.

- Was ist der Sinn eines solchen Intervalls, welche Information gewinnt man dadurch? Formulieren Sie mit eigenen Worten.

Bei der Bestimmung eines solchen Intervalls soll mit dem Werkzeug der Binomialverteilung gearbeitet werden.

- Vergegenwärtigen Sie sich die grundlegenden Eigenschaften der Binomialverteilung.
- Verwenden Sie die Daten aus der Stichprobe des Arbeitsblocks 1, um die Parameter einer Binomialverteilung festzulegen. Interpretieren die so festgelegte Binomialverteilung im Zeitungskontext der Stichprobenziehung.

Benutzen Sie eine Simulationsdatei, um den Wahrscheinlichkeitsparameter p der Binomialverteilung gezielt zu variieren:

- Bestimmen Sie p so, dass nur noch 5 % der Verteilung über dem Textanteil liegen. Interpretieren Sie diesen p-Wert im Kontext.
- Bestimmen Sie p einmal so, dass nur noch 2,5 % der Verteilung über dem Textanteil liegen (p_u) und einmal so, dass 2,5 % der Verteilung unter dem Textanteil liegen (p_o).
- Interpretieren Sie das Intervall $[p_u ; p_o]$ in mathematischer und in kontextueller Hinsicht.

Vergleichen Sie anschließend das Simulationsergebnis mit dem Formelergebnis für ein Näherungsintervall (Schulbuch).

Die im zweiten Arbeitsblock vorgeschlagene Zugangsweise geht nicht den algorithmischen Weg über die Vorgabe und Anwendung der Formel. Sie spiegelt eine induktive Zugangsweise wider, bei der Schülerinnen und Schüler das ihnen zur Verfügung stehende mathematische Werkzeug der Binomialverteilung mit den zugrundeliegenden mathematischen Modellvoraussetzungen in einem Sachkontext kritisch reflektierend zur Anwendung bringen können. Diese Vorgehensweise folgt der Idee Vollraths (2003), Funktionen zur Umwelterschließung (was die mathematische Welt explizit einschließt) zu nutzen. Dazu gehören außer der Lösung selbst mathematische Begründungen für die Vorgehensweise, z. B. die Legitimation mit der Binomialverteilung (Text vs. Nicht-Text) zu arbeiten oder die Modellannahme, dass die Stichproben durch die gleiche Ziehungsvorschrift zustande gekommen sind und zumindest prinzipiell beliebig oft wiederholbar sind. Diese Vorgehensweise folgt den Überlegungen von Vehling (2011) und von Eichler & Vogel (2011; 2012).

Mit einer entsprechenden Rechner-Datei (Fathom, Excel, Geogebra oder TI-Nspire), die je nach Kenntnisstand der Klasse und der zur Verfügung stehenden Unterrichtszeit entweder durch die Schülerinnen und Schüler selbst erstellt oder von der Lehrkraft vorgegeben wird, kann die Binomialverteilung für ein 95 %-Konfidenzintervall durch gezieltes Ausprobieren mit einem Schieberegler für den Wahrscheinlichkeitsparameter p so angepasst werden, bis

- nur noch 2,5 % der von der Binomialverteilung eingeschlossenen Fläche oberhalb des ermittelten Textanteils h_a liegen (p_u)
- nur noch 2,5 % der von der Binomialverteilung eingeschlossenen Fläche unterhalb des ermittelten Textanteils h_a liegen (p_o)

wobei p_u und p_o die Grenzen des Konfidenzintervalls kennzeichnen.

Arbeitsblock 3 (Schätzmethode vertiefen):

- Legen Sie die Anteile p_u und p_o symmetrisch um den ermittelten Textanteil. Wie lässt sich dies mathematisch als sinnvoll rechtfertigen?
- Untersuchen Sie mit der Simulationsdatei
  - den Einfluss unterschiedlicher Konfidenzniveaus (z. B. 99 %, 95 %, 80 %) bei gleichbleibender Stichprobengröße und
  - den Einfluss unterschiedlicher Stichprobengrößen bei gleichbleibendem Konfidenzniveau

  auf die Intervallbreite: Was bedeutet dies für den Informationsgehalt?

Interpretieren Sie Ihr Arbeitsergebnis bezogen auf die Frage nach dem Textanteil in der gesamten Zeitung.

Formulieren Sie ein Fazit.

Wenn die Schülerinnen und Schüler Erfahrungen mit der Entwicklung der Schätzmethode gesammelt haben, dann kann man im Sinne einer effizienteren Vorgehensweise noch einen Schritt weiter gehen.

Da für hinreichend großes n die Verteilungen nahezu symmetrisch sind, können die Grenzen p_u und p_o per Konvention auch symmetrisch um den empirisch ermittelten Textanteil h_a gelegt werden. Das Konfidenzintervall ergibt sich dann zu:

$[h_a - (h_a - p_u);\ h_a + (h_a - p_u)] = [p_u;\ 2 \cdot h_a - p_u]$. Auf diese Weise lässt sich die Bestimmung des Konfidenzintervalls mit nur einem Schieberegler für p_u in einem Vorgang erledigen.

Im vorliegenden Fall der oben genannten Daten ergibt sich so das ermittelte, in Abbildung 3 gezeigte Konfidenzintervall von [0,460; 0,534] für den relativen Textanteil $h_a = 0{,}497$. Es zeigt sich, dass in ca. 4,9 % von 100 vergleichbaren Fällen die unbekannte wahre Wahrscheinlichkeit außerhalb des Intervalls liegen würde. Die Interpretation dieses Sachverhalts ist ebenso Aufgabe für die Schülerinnen und Schüler wie die Bestimmung des Konfidenzintervalls selbst.

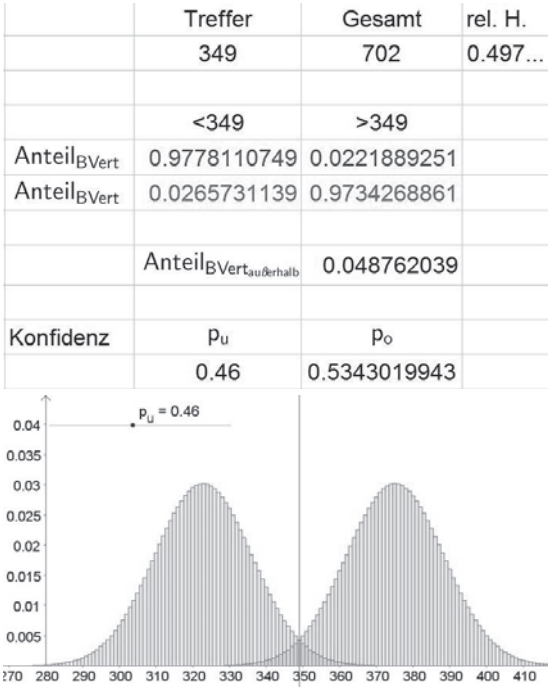


Abb. 3: Rechnergestützte Bestimmung eines Konfidenzintervalls

Wenn man von der annähernd symmetrischen Verteilung einer binomialverteilten Zufallsgröße ausgehen kann (als Faustformel gilt dies bei einer Varianz $V(X) = n \cdot p \cdot (1-p) > 9$), lassen sich die bekannten Sigma-Regeln für die Normalverteilung als Abschätzung für die Binomialverteilung nutzen. Als Näherungsintervalle ergeben sich:

$$h_n \pm \frac{1{,}96\sqrt{h_n(1-h_n)}}{\sqrt{n}} \text{ oder } h_n \pm \frac{1}{\sqrt{n}}$$

(da die Wurzel im Zähler mit $0 \le h_n \le 1$ maximal den Wert 0,5 annimmt, was bei Multiplikation mit dem Faktor 1,96 nach oben abgeschätzt zu einem Zählerwert von 1 führt). Das Konfidenzintervall ergibt sich als Lösung einer (umständlich zu lösenden) Ungleichung

$$|p - h_n| \le \frac{1{,}96\sqrt{p(1-p)}}{\sqrt{n}}.$$

Wenn sich die Schülerinnen und Schüler im vorliegenden Fall der Mühe unterziehen, das Konfidenzintervall über die Lösung dieser Ungleichung zu ermitteln, werden sie feststellen, dass Näherung und exakte Berechnung bis auf drei Nachkommastellen übereinstimmen.

Neben dieser Feststellung ist aber in didaktischer Hinsicht insbesondere wichtig, dass die Schülerinnen und Schüler mit der Rechner-Datei die Möglichkeit haben, den Einfluss unterschiedlicher Konfidenzniveaus und unterschiedlicher Stichprobengrößen auf Intervallbreite und Informationsgehalt zu untersuchen, um so die Grundidee eines Konfidenzintervalls besser verstehen zu lernen:

- Vergrößern sie das Konfidenzniveau auf 99 %, vergrößert sich auch das Konfidenzintervall auf [0,448; 0,546].
- Bei der Betrachtung einer kleineren Stichprobe (etwa StP_01 anstelle der aggregierten Stichprobe) ergibt sich bei gleichbleibendem Konfidenzniveau von 95 % ebenfalls ein vergrößertes Konfidenzintervall von [0,323; 0,525].

Aus weiteren Versuchen des systematischen Probierens werden die Schülerinnen und Schüler zum einen feststellen können, dass bei Erhöhung des Konfidenzniveaus das Konfidenzintervall (bei gleichbleibender Stichprobengröße) größer wird, dafür aber der Informationsgehalt sinkt. Andererseits wird das Konfidenzintervall bei Erhöhung der Stichprobengröße (unter Voraussetzung gleicher Modalitäten der Stichprobenziehung und gleichbleibendem Konfidenzniveau) kleiner, der Informationsgehalt steigt. Im Kontext des Zeitungsbeispiels heißt dies, dass man die Stichprobe erhöhen (Anzahl der Lochungen) oder das Konfidenzniveau (z. B. 90 % oder 80 %) senken muss, um ein kleineres Konfidenzintervall zu erhalten und dadurch den Informationsgehalt zu erhöhen. Bei der Erhöhung der Stichprobe kann man sich allerdings die Erhöhung des Informationsgehalts nur durch erhebliche Anstrengung erkaufen, da der Faktor m, mit dem man die Stichprobe erhöht nur mit der Wurzel ($\sqrt{m}$) in die Verkleinerung des Konfidenzintervalls eingeht.

Arbeitsblock 4 (Schätzmethode anwenden):

- Wenden Sie die erarbeitete Schätzmethode für den Textanteil in einer anderen Zeitschrift und/ oder eine Zeitung an.
- Vergleichen Sie die verschiedenen Schätzungen. Welche Schlussfolgerungen lassen sich ziehen?

Beim Vergleich sollten sich beispielsweise bildlastige Illustrierten und textlastige Zeitungen voneinander unterscheiden lassen. Bei einer Stichprobenziehung aus einer Bildillustrierten, die entsprechend der Vorgehensweise im Arbeitsblock 1 vorgenommen wurde ergab sich folgender Datensatz für eine Januarausgabe der Bildillustrierten View im Jahr 2014 (Angaben: Stichprobennummer: Anzahl der Löcher – Anzahl Textlöcher – rel. Häufigkeit):

StP_01: 86-8-0,093	StP_02: 92-11-0,120
StP_03: 96-5-0,052	StP_04: 96-26-0,271
StP_05: 92-4-0,043	StP_06: 84-8-0,095
StP_07: 90-0-0,000	StP_08: 86-12-0,140

Daraus ergibt sich für die entsprechend aggregierte Stichprobe von 74 Textlöchern (bei einer Gesamtzahl von 722 Löchern und einem relativen Anteil von 0,102 ein Konfidenzintervall von [0,08; 0,125] für den relativen Textanteil. Interpretieren die Schülerinnen und Schüler diese Ergebnisse auf dem Hintergrund der einführenden Leseranfrage, lässt sich festhalten, dass die gedruckte Informationsweitergabe verschiedenen Bewertungskriterien unterliegt. Der Ausdruck „Blei-Wüste“ zeigt, dass der angemessene Anteil einer Bebilderung den Fluss und die Motivation des Lesevorgangs unterstützen soll. Die für die Tageszeitung und Bildillustrierten ermittelten Werte scheinen in ihrer Größenordnung durchaus glaubhaft (was sich auch durch entsprechende Nachfragen bei den Verlagen bestätigt). So legitimiert sich die mathematische Zugangsweise auch im Hinblick auf die nachträgliche Bestätigung aus der „Real-Welt“. Dieser Umstand birgt in didaktischer Hinsicht großes Potential, die Relevanz und Belastbarkeit des eigenen mathematischen Tuns für die Schülerinnen und Schüler erfahrbar werden zu lassen. So kann in pädagogischer Hinsicht ein Beitrag dazu geleistet werden, das Zutrauen der Schülerinnen und Schüler in die eigenen mathematischen Fähigkeiten zu stärken.

3 Kriterien zur Beurteilung und Generierung von Beispielen

Beispiele stehen als musterhafter Einzelfall für etwas Allgemeineres dahinter und weisen (im Idealfall) prototypisch auf den oder zumindest einen wesentlichen Aspekt dessen hin, worum es im Kern gehen soll. Im Zusammenhang von Fragen des Lehren und Lernens wusste bereits Seneca über das Potential von Beispielen festzuhalten: „longum iter est per praecepta, breve et efficax per exempla.“ [„Durch Lehren ist der Weg lang, durch Beispiele ist er kurz und wirksam.“, Übersetzung der Autoren] Hinter dieser Auffassung steckt die Einsicht bzw. die Erfahrung, dass das hinter einem Beispiel stehende Allgemeinere in inhaltlicher Hinsicht umfassender und in struktureller Hinsicht tiefer und dichter vernetzt ist und eine größere Abstraktion zulässt. Es erschließt sich dem Experten, der bereits kundig ist und der sich ein abstraktes Bild des Allgemeineren aus einer Vielzahl vorausgegangener (in der konkreten Erinnerung vermutlich meist „gesunkener“) Beispiele synthetisieren kann – es ist für den Experten gewissermaßen konkret genug geworden. Novizen sind demgegenüber auf konkrete Beispiele angewiesen, die ihnen aus ihrer bisherigen Denk- und Erfahrungswelt heraus einen konkreten Zugang zu den hinter den Beispielen stehenden allgemeineren Ideen eröffnen.

An dem einführenden Beispiel des „Zeitungslochens“ als handlungsorientierte unterrichtliche Umsetzung zum Thema Stichprobenziehung und Konfidenzintervall lassen sich didaktische Kriterien festmachen, die bei der Gestaltung und Beurteilung beispielgebundener Zugänge relevant sind:

Schülerwelt-Bezug:

Entsprechende Beispiele sollten den Schülerinnen und Schülern einen möglichst unmittelbaren und konkreten, im wörtlichen Sinn „greifbaren“ Zugang zur zentralen Fragestellung eröffnen. Inhaltlich stellen solche Beispielaufgaben, die Freudenthals paradigmatischen Grundgedanken von „Beziehungshaltigkeit“ (Freudenthal, 1973, S. 75 ff.) folgen, Problemstellungen dar, die den Schülerinnen und Schülern Lerngelegenheiten bieten sollen, „Phänomene ihrer natürlichen, technischen, geistigen und sozialen Umwelt“ (Klieme, Neubrand, & Lüdtke, 2001, S. 142) zu erschließen.

Wenn in diesem Zusammenhang von Problemauthentizität gesprochen wird, ist damit weder gemeint, dass nur wirkliche Probleme der natürlichen, technischen und gesellschaftlichen Umwelt den Schülerinnen und Schülern zur Bearbeitung vorgesetzt werden dürften. Dies müsste schon allein an der Tatsache scheitern, dass die meisten Probleme mit diesem Anspruch für die unterrichtliche Aufbereitung schlicht zu komplex und zu zeitaufwändig wären (gleichwohl gibt es geeignete Beispiele, die dann auch Eingang in den Unterricht finden sollten, vgl. Abschnitt 4). Auch eine in der Unterrichtspraxis immer wieder zu beobachtende motivationale „Anbiederung“ an die vermeintliche Erlebnis- und Erfahrungswelt der Schülerinnen und Schüler durch entsprechend kontextuell adaptierte Beispiele funktioniert in der Regel nicht, ohnehin nicht für alle Schülerinnen und Schüler einer Klasse. Es geht schlicht um die unmittelbare intellektuelle Fassbarkeit der dem Beispiel zugrundeliegenden

Idee: Im Textlochen-Beispiel wird ein Unterrichtspraktiker zweifellos davon ausgehen, dass Schülerinnen und Schüler der gymnasialen Oberstufe in der Lage sind, die Idee der Stichprobenziehung hinter dem Beispiel zu sehen. Eine erste Schätzung ist nicht nur in statistischer Hinsicht wichtig, sie liefert auch in didaktischer Hinsicht das Motiv zur Überprüfung, sie löst das fragende Interesse aus. Die praktische Zugangsmöglichkeit der verwendeten Materialien (Locher, Zeitung und Rechner), kann dann bei allen Folgeschritten der Datenerhebung (Lochen als Stichprobenziehung), -aufbereitung (von der Punktschätzung zur Intervallschätzung) und -bewertung (Interpretation und Vergleich mit der vorausgehenden Schätzung) immer wieder eine praktische Referenz liefern: Einen „sicheren Hafen“ als Rückzugsmöglichkeit ins Konkrete, wenn die theoretischen Überlegungen schwer werden. Dennoch können auch Beispiele, die bewusst für den Unterricht in einen für Schülerinnen und Schüler fassbaren Kontext gesetzt werden, durchaus Parallelität zu realen Beispielen aufweisen: Genauso, wie eine beliebige Eigenschaft etwa der Deutschen Bevölkerung prinzipiell bekannt sein kann, ist auch der Textanteil einer Zeitung bekannt. Da aber die exakte Erhebung praktisch nicht möglich scheint, behilft man sich mit einer möglichst gut konstruierten Stichprobe, um den wahren Anteil abzuschätzen.

Mathematik-Bezug:

In mathematischer Hinsicht ist von Bedeutsamkeit, dass das mathematische Konzept oder Verfahren, das zu erlernen im Beispiel angebahnt werden soll, sich in der Modellierung des Beispielkontextes im Kern abbilden und glaubhaft rechtfertigen lässt. So liegt bereits der Auswahl einer konkreten Tageszeitung, deren Seiten gelocht werden, die Annahme zugrunde, dass diese repräsentativ für die Grundgesamtheit aller dieser Tageszeitungen ist (zumindest für die Jahrgänge, die dem gleichen layouttechnischen Verfahren unterworfen sind). Zudem wird dem Vorgang des Lochens als zufälliger Ziehungsvorgang interpretiert implizit unterstellt, unendlich oft wiederholbar zu sein. Darin steckt die Annahme und Begründung für die Beurteilung der Stichprobe als repräsentativ für die Grundgesamtheit gelten zu können. Dabei wird mit dem Maß der Anzahl von ausgestanzten Kreisflächen gemessen, auch dies eine modellierungstechnische Entscheidung, zu der es durchaus Alternativen gibt. Dem Lochungsverfahren liegen folgende mathematische Überlegungen zugrunde: Das n-Tupel von Zufallsgrößen $(X_1, X_2, \ldots, X_n)$ beschreibt die mathematische Stichprobe vom Umfang n, die Zufallsgröße X_i beschreibt die zufällige Merkmalsausprägung des i-ten Loches. Es wird die Annahme zugrunde gelegt, dass alle X_i der gleichen Ziehungsvorschrift unterliegen und somit die gleiche Verteilung F_X besitzen. Eine konkrete Stichprobe $(x_1, x_2, \ldots, x_n)$ mit x_i als Merkmalsausprägung des i-ten Loches (1 für Text, 0 für kein Text) ist eine empirische Realisierung der mathematischen Stichprobe, also z. B. im Fall von der o. g. StP_01 bei der Bildillustrierten View ein 86er Tupel von Nullen und Einsen mit insgesamt 8 Einsen.

Mathematische Überlegungen dieser Art werden im Allgemeinen nicht mit den Schülerinnen und Schülern im Vorhinein diskutiert werden (können), ohnehin nicht bei einer beispielgebundenen induktiven Zugangsweise zu einem mathematischen Konzept, wie hier im oben geschilderten Zugang zum Konfidenzintervall über die Grenzbetrachtungen entsprechend überlappender Binomialverteilungen im Zeitungsbeispiel. Die Lehrkraft muss sich aber im Klaren über den mathematischen Hintergrund des gewählten Zugangsbeispiels sein, um beurteilen zu können, ob sich die von den Schülerinnen und Schülern zu leistende Modellierung mathematisch rechtfertigen lässt und auf eine adäquate mathematische Begrifflichkeit führen kann.

Unterrichtspraktische Umsetzung:

Dieser Aspekt ist ganz von der Pragmatik unterrichtlicher Umsetzbarkeit geprägt: Die schönste, schülernahste und mathematisch gehaltvollste Beispielidee nützt nichts, wenn sie sich nicht mit einem angemessenen Aufwand an zeitlichen und sächlichen Ressourcen im Unterricht realisieren lässt. Unterrichtspraktiker, die sich dem engen schulischen Zeitkorsett vor Ort ausgesetzt sehen, haben einen Begriff für diesen Umstand: Feiertagsstunden. Es mag in gewisser Weise das Schicksal eines datenorientierten Stochastikunterrichts sein, einerseits zwar durch die Daten vermittelt authentische und gehaltvolle Modellierungsaktivitäten initiieren zu können (dies sind tatsächlich Feiertagsstunden in bestem Sinne) wie dies kaum in anderen Bereichen der Mathematik möglich ist, andererseits dies aber mit einem erhöhten unterrichtlichen Zeitbedarf – etwa gegenüber einer kalkülorientierten unterrichtlichen Abhandlung der kombinatorischen Grundfiguren – zu bezahlen (zumindest in der häufig geäußerten Wahrnehmung betreffender Lehrkräfte). Umso wichtiger ist, dass sich der materielle Aufwand in überschaubaren Grenzen hält und auch der oftmals als vermeintlich unmathematisch betrachtete Teil, die Datenerhebung (so denn eine geplant oder vonnöten ist) nicht zuviel Zeit in Anspruch nimmt.

Im vorliegenden Beispiel des Zeitungslochens sind beide Kriterien, sparsamer sächlicher und zeitlicher Ressourcenverbrauch, gegeben: Locher gibt es genügend in jeder Schule, Zeitungen oder Illustrierten sind problemlos zu beschaffen (hier können die Schülerinnen und Schüler eigene, für sie interessante Zeitschriften mitbringen) und die Datenerhebung ist rasch gemacht. Nach unseren Erfahrungen ist bei Vorgabe geeigneter Rechner-Dateien und Arbeitsblätter[1] die unterrichtliche Umsetzung mitsamt der Datenerhebung in einer Doppelstunde machbar, was freilich nicht heißen soll, dass damit das Thema Konfidenzintervalle abschließend behandelt worden wäre.

Es mögen sicherlich noch weitere und individuellere Faktoren der Unterrichtsplanung bei der didaktischen Bewertung eines beispielgebundenen Zugangs eine Rolle spielen: Wir haben an dieser Stelle diejenigen herausgehoben, die sich aus der strukturellen Zusammenschau verschiedener didaktischer Modelle (etwa der bildungstheoretischen Didaktik nach Klafki, 1957 oder der lehrtheoretischen Didaktik nach Schulz, 1970) unseres Erachtens am unmittelbarsten ergeben.

4 Beispiele schaffen Zugänge

Allgemein bezogen auf den Mathematikunterricht erfolgt ein solcher beispielbezogener Zugang in der Regel über das eingesetzte Aufgabenmaterial. In den Jahren seit den ersten Veröffentlichungen von TIMSS und PISA wurde viel an mathematikdidaktischer Entwicklungs- und Forschungsarbeit in diesem Bereich geleistet (vgl. z. B. Herget, 2000; Büchter & Leuders, 2005; Bruder, Leuders & Büchter, 2008). Zusammenfassend lässt sich festhalten, dass Ausgangspunkt und paradigmatische Ausrichtung der Bemühungen um gute Aufgaben ein guter Unterricht ist (vgl. Leuders, 2001, S. 94 ff.) und solche Aufgaben, entsprechend unterrichtlich ein- bzw. umgesetzt, aktiv-entdeckendes Lernen ermöglichen, ein stimmiges Bild von Mathematik und ihren Anwendungen zeichnen, konkurrierende Lösungsansätze zulassen und Erfahrungen für mathematische Begriffsbildungen bieten (vgl. BLK, 1997, S. 14 ff; Büchter & Leuders, 2005, S. 13).

In spezifischem Fokus auf einen datenorientierten Stochastikunterricht stehen insbesondere Modellierungsbeispiele im Mittelpunkt des Interesses: Daten sind als Kontextzahlen immer mit einem situativen Problemhintergrund verbunden, zu dem die Daten quantifizierende Informationen bereit stellen. In didaktischer Hinsicht lassen sich entsprechende Beispielaufgaben unterscheiden (vgl. Eichler & Vogel, 2013, S. 140 ff.):

Beispiele zur „realen Realität“, bei deren Bearbeitung die Schülerinnen und Schüler in die Lage versetzt werden zu erfahren, dass sie mit den ihnen zur Verfügung stehenden mathematischen Mitteln echte (im Sinne von nicht von vorneherein für den Unterricht konstruierten) Fragen der sie umgebenden technischen, sozialen und natürlichen Umwelt zumindest ein Stück weit beantworten können. Hierzu zählen etwa gesellschaftspolitische Fragen, wie z. B. die mit dem Ärzteprotest verbundenen politischen Diskussionen, ob Ärzte in Deutschland zu viel oder zu wenig verdienen (vgl. Eichler & Vogel, 2013; S. 41 ff.). Beispiele aus diesem Bereich dienen der rekonstruktiven Datenanalyse (ibid., S. 140), durch die die Schülerinnen und Schüler erfahren sollen, mit welchen stochastischen Mitteln in der Realität argumentiert werden kann bzw. auch tatsächlich wird. Beispiele der rekonstruktiven Datenanalyse sollen unmittelbar dem Ziel dienen, die Schülerinnen und Schüler zu einer kritischen und mündigen Teilhabe am Leben einer demokratischen aufgeklärten Gesellschaft zu befähigen.

Beispiele zu konstruierten realen Situationen beziehen sich nicht auf die „großen Fragen der Gesellschaft“, sondern auf „kleinere“, durchaus auch künstlich generierte, aber dennoch reale Phänomene. Hierbei kann es sich um Fragen handeln, die für die Schülerinnen und Schüler unmittelbar von Belang sind (vor dem Computer verbrachte Zeit, vgl. Eichler, 2009), aber auch um Analysen von im Internet erhältlichen Daten zu Sportgeschehnissen wie z. B. Skispringen (Eichler & Vogel, 2013, S. 89 ff.), Fragen nach der Sicherheitsgrenze für einen Nicht-Abstieg aus der Fussball-Bundesliga (ibid., S. 48 ff.) oder Fragen der physikalischen Umwelterschließung, wie z. B. die kritische Betrachtung verfügbarer meteorologischer Daten (ibid., S. 21 ff.), der Frage nach der Schallgeschwindigkeit (ibid., S. 103 ff.) oder dem weltweiten CO_2-Anstieg (ibid., S. 107 ff.). Aber auch bei der Arbeit mit „Mickey-Mouse“-Daten (vgl. Engel, 2007, S. 16), welche nicht unmittelbar die Eigenschaften der Schülerinnen und Schüler betreffen und auch inhaltlich (in der Regel) nicht weiter ernst zu nehmen sind, wie z. B. der Sprungweitenvergleich von großen und kleinen Papierfröschen (Eichler & Vogel, 2013, S. 7 ff.) oder die Frage nach der Farbverteilung von Schokolinsen (Engel & Vogel, 2005; Eichler & Vogel, 2013, S. 31 ff.), können die Schülerinnen und Schüler den gesamten Modellierungskreislauf in Form einer datenanalytischen Modellierung (Eichler & Vogel, 2013a) durchlaufen: Problemstellung – Planung und Durchführung der Datenerhebung – Auswertung – Interpretation und Schlussfolgerungen (vgl. Biehler & Hartung, 2006, S. 53).

Beispiele zur Analyse konstruierter Daten dienen dazu, die Reichweite und Grenzen statistischer Begriffe und Methoden für die Schülerinnen und Schüler erfahrbar werden zu lassen. Die Daten sind in diesem Fall fiktiv und offensichtlich, auch für die Schülerinnen und Schüler, gezielt so konstruiert, dass mathematische Eigenschaften wie z. B. der Robustheit des Medians (gegenüber dem arithmetischen Mittel) anhand von sechs fiktiven Ärzteeinkommensdaten (Eichler & Vogel, 2013, S. 43) oder unterschiedliche mathematische Konzepte zur Beschreibung einer „Mitte“ anhand von konstruierten Firmengehältern (ibid., S. 45 f.) deutlich heraustreten. Hier wird der situative Problemhintergrund in seiner Bedeutsamkeit bewusst soweit reduziert, dass lediglich ein inhaltlicher Rahmen für die fokussierte Betrachtung statistischer Begriffe und Methoden verbleibt.

Im Hinblick auf die didaktische Verortung im Unterrichtsgeschehen steht bei der unterrichtlichen Analyse konstruierter Daten der Zugang zum sachverständigen mathematischen Werkzeuggebrauch im Vordergrund der unterrichtlichen Bemühungen. Dagegen eignen sich die ersten beiden Beispieltypen eher, um thematische Zugänge im Sinne eines anwendungsbezogenen Mathematikunterrichts zu schaffen: Bei Beispielen zur „realen Realität“ und zu konstruierten realen Situationen steht der realweltliche Erkenntnisgewinn im Mittelpunkt des unterrichtlichen Interesses. Die Schülerinnen und Schüler wenden anhand ausgewählter Beispiele statistische Arbeitsweisen und Verfahren an, um erklärende oder zumindest erhellende Antworten auf eine kontextuelle Ausgangsfrage zu erhalten.

5 Alles neu …?

Über die Eignung der Beispiele hat die Lehrkraft nach verschiedenen Kriterien zu entscheiden. Die in Abschnitt 3 genannten Kriterien sind aus einer unterrichtspraktischen Sichtweise heraus entwickelt dargestellt. Das theoretische Fundament dazu lässt sich unserer Auffassung nach überzeugend im Prinzip des *Exemplarischen* der bildungstheoretischen Didaktik (z. B. Klafki, 1957) verorten: Als *exemplarisch* ist nicht das Beispiel zu bezeichnen, das lediglich unter dem Kriterium der Vermeidung von Stofffülle gewählt wird. Bildend im bildungstheoretischen Sinn sind Beispiele dann, wenn sie *„elementar im Hinblick auf die Sache* (im Besonderen ein Allgemeines zeigen) und wenn sie *fundamental im Hinblick auf die Schüler* (Grunderfahrungen und grundlegende Einsichten vermitteln)“ (Jank & Meyer, 1991, S. 146) sind, als pädagogisch-exemplarisch können sie dann bezeichnet werden, wenn „sie Fundamentales oder Elementares aufzuschließen vermögen“ (Klafki, 1961, S. 191). Die Zitate und die genannten Quellen zeigen, dass diese Einsichten nicht neu, sondern im Gegenteil einer längeren Tradition bildungstheoretischer Überlegungen folgen und gerade deshalb lohnt es sich, dies im aktuellen Diskurs der heutigen, zuweilen sehr „bildungsmetrischen“ Zeit wieder einmal in Erinnerung zu rufen.

Anmerkung

1 Entsprechende Arbeitsblätter und Rechner-Dateien können von den Autoren für rein unterrichtliche Zwecke bezogen werden.

Literatur

Biehler, R. & Hartung, R. (2006). Leitidee Daten und Zufall. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Eds.), *Bildungsstandards Mathematik konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (S. 51–80). Berlin: Cornelsen Scriptor.

BLK – Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung (Hrsg.). (1997). *Gutachten zur Vorbereitung des Programms zur Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts.* Bonn.

Bruder, R., Leuders, T. & Büchter, A. (Hrsg.) (2008). *Mathematikunterricht entwickeln. Bausteine für ein kompetenzorientiertes Unterrichten.* Berlin: Cornelsen Scriptor.

Büchter, A. & Leuders, T. (2005). *Mathematikaufgaben selbst entwickeln. Lernen fördern – Leistung überprüfen.* Berlin: Cornelsen Scriptor.

Eichler, A. (2009). Zahlen aufräumen – Daten verstehen (Basisartikel). *PM – Praxis der Mathematik in der Schule, 51*(26), 1–7.

Eichler, A. & Vogel, M. (2013). *Die Leitidee Daten und Zufall.* Wiesbaden: Springer Spektrum.

Eichler, A. & Vogel, M. (2013a). Daten- und Wahrscheinlichkeitsanalyse als Modellierung. In R. Borromeo Ferri, G. Greefrath & G. Kaiser (Hrsg.), *Mathematisches Modellieren für Schule und Hochschule* (S. 163–180). Wiesbaden: Springer Spektrum.

Eichler, A. & Vogel, M. (2012). Stochastik – fit für die Zukunft (Basisartikel). *PM – Praxis der Mathematik in der Schule, 54*(48), S. 2–9.

Eichler, A. & Vogel, M. (2011). *Leitfaden Stochastik.* Wiesbaden: Vieweg+Teubner.

Engel, J. (2007). Daten im Mathematikunterricht: Wozu? Welche? Woher? *MU – Der Mathematikunterricht, 3*, 12–22.

Engel, J. & Vogel, M. (2005). Von M&Ms und bevorzugten Farben: ein handlungsorientierter Unterrichtsvorschlag zur Leitidee Daten & Zufall in der Sekundarstufe I. *Stochastik in der Schule, 25*(2), 11–18.

Freudenthal, H. (1973). *Mathematik als pädagogische Aufgabe.* Stuttgart: Klett.

Herget, W. (Hrsg.) (2000). Aufgaben öffnen. *mathematik lehren, 100.*

Jank, W. & Meyer, H. (1991). *Didaktische Modelle.* Frankfurt am Main: Cornelsen Scriptor.

Klafki, W. (1957). *Das pädagogische Problem des Elementaren und die Theorie der kategorialen Bildung.* Weinheim: Beltz.

Klafki, W. (1961). Das Elementare, Fundamentale, Exemplarische. In H.-H. Groothoff & M. Stallmann (Hrsg.), *Pädagogisches Lexikon* (S. 189–194). Stuttgart: Kreuz-Verlag.

Klieme, E., Neubrand, M. & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In: J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139–190). Opladen: Leske+Budrich.

Leuders, T. (2001). *Qualität im Mathematikunterricht der Sekundarstufe I und II.* Berlin: Cornelsen Scriptor.

Schulz, W. (1970). Aufgaben der Didaktik. Eine Darstellung aus lehrtheoretischer Sicht. In D. C. Kochan (Hrsg.), *Allgemeine Didaktik – Fachdidaktik – Fachwissenschaft. Ausgewählte Beiträge aus den Jahren 1953 bis 1969* (S. 403–440). Darmstadt: Wissenschaftliche Buchgesellschaft.

Seneca, L. A. Epistularum moralium ad Lucilium, *liber primus*, epistula 6, URL: http://www.thelatinlibrary.com/sen/seneca.ep1.shtml (Stand: 20.6.2014).

Vehling, R. (2011). Mit Simulationen zum Konfidenzintervall. *PM – Praxis der Mathematik in der Schule, 53* (39), 25–29.

Vollrath, H.-J. (2003). *Algebra in der Sekundarstufe.* Heidelberg: Spektrum

Anschrift der Verfasser

Markus Vogel
Pädagogische Hochschule Heidelberg
Im Neuenheimer Feld 561
69120 Heidelberg
vogel@ph-heidelberg.de

Andreas Eichler
Institut für Mathematik
Universität Kassel
Gebäude AVZ III, Raum 2435
Heinrich-Plett-Straße 40
D-34132 Kassel
eichler@mathematik.uni-kassel.de

Inferenzstatistik per Simulation: Bootstrap-Konfidenzintervalle in der Sekundarstufe II mit Excel

Andreas Kaufmann, Echallens, Joachim Engel, Ludwigsburg

Zusammenfassung: *Dieser Aufsatz enthält einen Vorschlag für die Einführung von Bootstrap-Konfidenzintervallen in der Sekundarstufe II. Zuerst wird der Nutzen dieser Methoden für die angewandte Statistik und für den Mathematikunterricht erläutert. Anschließend wird das Verfahren an zwei Beispielen präsentiert, seine Eigenschaften werden anschaulich begründet und Fehlerquellen und Konvergenzfragen werden diskutiert. Schließlich werden Vorschläge zum Erweitern und Verfeinern präsentiert.*

1 Einleitung

Erheben von Daten, Erstellen von Grafiken wie z. B. Histogramme, Streudiagramme oder Errechnen numerischer Werte wie Mittelwerte, Mediane oder Streumaße sind sehr konkrete Operationen. Dagegen sind Konzepte, die auf der Zufallsvariabilität dieser Statistiken beruhen wie Stichprobenverteilungen, Standardfehler, Konfidenzintervalle, Hypothesentests und P-Werte abstrakt und schwer für Schüler zu verstehen. Bootstrap-Methoden nehmen die konkreten Konzepte, mit denen Schüler vom Arbeiten mit Daten vertraut sind, und wenden sie zur Herleitung von Stichprobenverteilungen an. Der Einsatz geeigneter Software hilft nicht nur den für Simulationen nötigen hohen Rechenaufwand zu bewältigen, sondern unterstützt vor allem auch konzeptionelles Verstehen. Wir illustrieren die Bootstrap-Methode zur Herleitung von (approximativen) Konfidenzintervallen für Mittelwerte und Korrelationskoeffizienten.

2 Bootstrap: Recycling von Daten

Die grundlegende Idee des Bootstraps ist wie folgt: Angenommen, man könnte beliebig oft Stichproben vom vorgegebenen Umfang *n* aus der Population ziehen und ließe sich dann aus jeder Stichprobe die Wer-

te der gewünschten Statistik (Mittelwert, Median, Varianz etc.) errechnen. Dann ließe sich aus zahlreichen solchen Stichprobenstatistiken die (empirische) Stichprobenverteilung erstellen. Nun kosten bei konkreten Anwendungen Stichproben Ressourcen (Zeit, Geld etc.) und das Ziehen von sehr vielen Stichproben gegebenen Umfangs ist aus praktischen Gründen indiskutabel. Bei statistischen Untersuchungen ist in der Regel die Population nicht verfügbar, sondern man hat nur eine Stichprobe. Hier setzt die Bootstrap-Idee an: Anstelle aus der nicht-verfügbaren Population werden weitere Stichproben – Bootstrap-Stichproben – aus der vorliegenden Stichprobe gezogen. Durch dieses Recycling der Daten (die Fachliteratur spricht von *Resampling*) können Schätzer zwar nicht verbessert werden, da der ursprünglich vorliegende Stichprobenumfang fest ist. Es ist jedoch möglich die Verteilung von Stichprobenstatistiken durch die Verteilung der Bootstrap-Stichproben zu approximieren. Diese Approximation ist umso besser, je mehr die vorliegende Stichprobe der Population ähnelt. Diese Idee geht auf Brad Efron (1979) zurück.

Bootstrap-Konfidenzintervalle erhält man dann, indem dann z. B. das 2,5 % und 97,5 % Perzentil der Bootstrap-Verteilung als Konfidenzgrenzen genommen wird. Bootstrap-Konfidenzintervalle sind viel flexibler als die klassischen Konfidenzintervalle, da die Berechnung analytisch unzugänglicher Stichprobenverteilungen – wie z. B. des Medians oder von Kennzahlen für robuste explorative Datenanalysen – beim Bootstrap durch Simulationen ersetzt werden, und weil man nicht auf Hypothesen hinsichtlich der Verteilung der Daten angewiesen ist (Engel & Grübel 2008). Die in den meisten Verfahren der klassischen Statistik vorausgesetzten Normalverteilungsannahmen sind hier nicht erforderlich. Das begründet die enorme Bedeutung dieser Methode für die moderne angewandte Statistik. Verschiedene Autoren (z. B. Cobb 2007, Engel 2010, Hesterberg 2014, Pfannkuch & Budgett 2014) erklären, dass der Bootstrap neben seiner Nützlichkeit als flexible Methode für viele Anwendungssituationen auch ein erhebliches didaktisches Potenzial besitzt. Die Rechenwege zur Gewinnung von Referenzverteilungen in der klassischen Inferenzstatistik sind in vielen Ausbildungsstufen – wie zum Beispiel in der Sekundarstufe II – nicht realisierbar. Simulationen bieten hier einen Ausweg, die formale Mathematik auf ein Minimum zu reduzieren und den Fokus auf konzeptionelles Verstehen zu richten.

Das zentrale Konzept der Inferenzstatistik ist die Stichprobenverteilung. Resampling Methoden erlauben dem Lernenden durch Simulationen zu erfahren, wie sich eine Statistik von Stichprobe zu Stichprobe unterscheidet. Anhand von Bootstrap-Konfidenzintervallen kann die beschreibende Statistik in der Sekundarstufe II zur Inferenzstatistik weitergeführt werden. Wenn Schüler die Idee des Bootstraps an einfachen Beispielen wie etwa dem Mittelwert verstanden haben, können sie das Verfahren leicht auf andere Kennwerte übertragen. Damit lassen sich dann vielfältige Projektarbeiten in der Sekundarstufe II durchführen.

Johnson (2001) zeigt in seinem Aufsatz, wie man Bootstrap-Konfidenzintervalle Studenten nahe bringen kann, die bereits mit der klassischen Inferenzstatistik vertraut sind. Christie (2004) erklärt wie man Bootstrap-Konfidenzintervalle mit Excel berechnen kann. Mit Excel lassen sich die einzelnen Schritte sowie das gesamte Verfahren gut visualisieren, das Programmieren ist relativ einfach und viele Schüler haben bereits Erfahrung mit diesem Programm. Im Rahmen dieses Aufsatzes gehen wir nicht näher auf die Gestaltung der Excel-Arbeitsblätter ein, die jedoch auf Anfrage von den Autoren zur Verfügung gestellt werden. Alternativ lässt sich der Bootstrap auch verständnisfördernd mit didaktisch konzipierter Stochastik-Software wie Fathom (Erickson 2001) oder Tinkerplots (Watson 2013) implementieren.

Auf begrifflicher Ebene verwenden wir den Begriff des *approximativen Modells der Population*, wenn wir von den Daten sprechen, dies um das Simulieren anhand der Daten zu veranschaulichen.

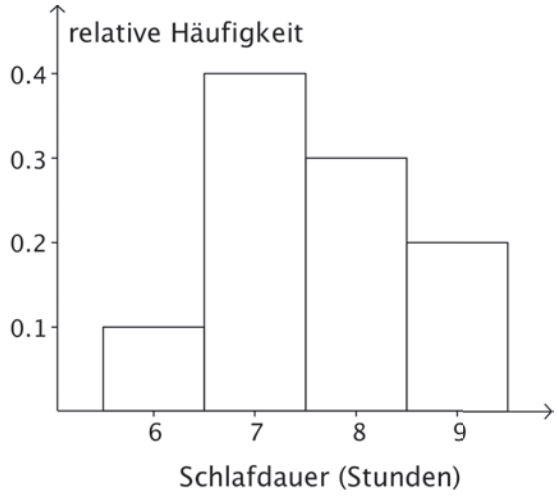
3 Bootstrap-Konfidenzintervalle für das arithmetische Mittel

Als einführendes Beispiel betrachten wir Konfidenzintervalle für den Mittelwert μ einer Population. Die Stichprobenverteilung von (empirischen) Mittelwerten ist annähernd symmetrisch verteilt (das sagt uns der zentrale Grenzwertsatz), wenn der Stichprobenumfang groß genug ist. Somit wird auch das Konfidenzintervall für den Populationsmittelwert annähernd symmetrisch sein. Im Folgenden soll die Idee von Bootstrap-Konfidenzintervallen illustriert werden. Um der Klarheit willen ist das konstruierte Beispiel bewusst einfach gewählt worden und mit einem wenig realistischen Kontext versehen. Es geht uns hierbei nicht um Authentizität einer Anwendung, sondern um ein Verständnis für das methodische Vorgehen beim Bootstrap. Die folgenden Überlegungen lassen sich leicht verallgemeinern auf Bootstrap-Konfidenzintervalle für Kennwerte, die durch Schätzer mit symmetrischer Stichprobenverteilung geschätzt werden.

3.1 Eine bekannte Population

Die Schulleitung einer Fachhochschule kennt die Schlafgewohnheiten aller Studenten, da bei der Immatrikulation alle Studenten in einem Fragebogen angeben mussten, wie lange sie täglich schlafen. Zunächst gehen wir also von der Annahme aus, dass die Population schon bekannt sei. Das ist zugegebenermaßen sehr unrealistisch, und – da die Population bekannt ist – erfordert es auch keinerlei Inferenzstatistik. Diese Annahme soll auch nur als illustrativer Zwischenschritt dienen. Durch Anwenden verschiedener Modelle (exaktes Modell und approximatives Modell) wird die Stimmigkeit der Verfahren dadurch plausibel, dass die ermittelten Werte zum Populationswert passen. Die Populationsdaten sind in Tabelle 1 zusammengefasst. Der Mittelwert der Population beträgt $\mu = 7{,}6$ Stunden.

Schlafdauer [in Stunden]	Relative Häufigkeit
6	10 %
7	40 %
8	30 %
9	20 %



Tab. 1: Relative Häufigkeiten der täglichen Schlafdauer der Population der Studenten der Fachhochschule und Histogramm.

3.2 Exaktes Modell der Population und 95 % Schwankungsbereich

Lena und Hasso studieren an der Fachhochschule. Im Rahmen einer Projektarbeit sollen sie unter anderem untersuchen wie lange ihre Mitstudenten täglich schlafen. Sie fragen gemäß Zufallsprinzip 20 Mitstudenten, wie lange sie täglich schlafen und berechnen das arithmetische Mittel. In welchem Bereich würde dieser Mittelwert in 95 % der „Fälle“ liegen, vorausgesetzt sie könnten die Befragung von 20 zufällig ausgewählten Mitstudenten beliebig oft wiederholen?

Diese Frage beantworten wir durch Simulieren von Umfragewerten: Als exaktes Modell für die Population betrachten wir die folgenden 10 Studenten und deren Schlafdauer (Tabelle 2). Dieses Modell spiegelt die Population repräsentativ wider, da die relativen Häufigkeiten dieser Stichprobe den relativen Häufigkeiten der Population gleichen.

Student	Schlafdauer [in Stunden]
1	6
2	7
3	7
4	7
5	7
6	8
7	8
8	8
9	9
10	9

Tab. 2: Exaktes Modell für die Population der Studenten der Fachhochschule.

Um eine Zufallsstichprobe mit 20 Umfragewerten zu erzeugen gehen wir so vor: Wir erzeugen eine Zufallszahl zwischen 1 und 10, sie entspricht dem Student aus dem Modell, der befragt wird, und wir schreiben die entsprechende Schlafdauer auf. Das machen wir 20 Mal und erhalten so 20 Umfragewerte, welche eine simulierte Zufallsstichprobe bilden (Spalte E in Tabelle 3), aus der sich ein simulierter Mittelwert errechnen lässt (Zelle E25 in Tabelle 3). Das wiederholen wir 1000 Mal und erhalten so 1000 simulierte Mittelwerte (Spalte G in Tabelle 3).

	A	B	C	D	E	F	G	H	I
1	Modell der Population			Simulierte Stichprobe			Wiederholte Mittelwerte		
2									
3	Student	Schlafdauer		Student	Schlafdauer		7.35		
4	1	6		9	9		7.7		
5	2	7		6	8		7.6	2.5% Quantil	
6	3	7		5	7		7.55		7.2
7	4	7		1	6		7.6		
8	5	7		9	9		7.8		
9	6	8		8	8		7.75		
10	7	8		1	6		7.4	97.5% Quantil	
11	8	8		8	8		7.05		8
12	9	9		4	7		7.5		
13	10	9		1	6		7.45		
14				1	6		7.45		
15	Mittelwert	7.6		10	9		7.6		
16				1	6		7.45		
17				7	8		7.85		
18				1	6		7.25		
19				8	8		7.65		
20				7	8		7.6		
21				7	8		7.75		
22				5	7		7.4		
23				4	7		7.5		
24							7.55		
25				Mittelwert	7.35		7.75		
26							7.7		

Tab. 3 : Anhand des exakten Modells der Population (Spalte B) wird eine Stichprobe simuliert (Spalte E) und deren Mittelwert berechnet (Zelle E16). Das wird 1000 Mal wiederholt und die Mittelwerte in Spalte G gespeichert.

Sei $q(0{,}025)$ das 2,5 % Perzentil dieser simulierten Mittelwerte und $q(0{,}975)$ das 97,5 % Perzentil. Bei

vielen Versuchen (z. B. bei 1000 Wiederholungen) liegt in ca. 95 % der Fälle der simulierte Mittelwert im Bereich $[q(0{,}025;\ q(0{,}975)] = [7{,}2;\ 8]$. Wir nennen dieses Intervall 95 %-Schwankungsbereich der Mittelwerte der Stichproben. Die halbe Breite des 95 %-Schwankungsbereiches ist $b = [q(0{,}975) - q(0{,}025)]/2 = 0{,}4$. Allgemein betrachtet gilt: Ist μ der Populationsmittelwert und $b = q(0{,}975) - \mu = \mu - q(0{,}025)$, dann ist der 95 %-Schwankungsbereich $[\mu - b, \mu + b]$.

3.3 Approximatives Modell der Population und Bootstrap-Konfidenzintervall für den Mittelwert

Lena und Hasso sind die Population nicht bekannt. Sie befragen gemäß Zufallsprinzip 20 Mitstudenten, wie lange sie täglich schlafen und erhalten 20 Umfragewerte (Tabelle 4), für die sich als arithmetisches Mittel $\bar{x} = 7{,}5$ errechnet. Lena sagt: „Hätten wir 20 andere Personen befragt, dann hätten wir einen anderen Mittelwert erhalten, vielleicht $\bar{x} = 7{,}1$“. Hasso sagt: „Oder vielleicht $\bar{x} = 7{,}8$. Wie sicher können wir uns bezüglich des Populationsmittelwerts sein?“ Gesucht ist ein um das Stichprobenmittel $\bar{x}$ symmetrisches Intervall, das den Populationsmittelwert μ in 95 % der Fälle enthält, d. h. in 100 hypothetisch durchgeführten Umfragen sollte das jeweilige Intervall in ca. 95 Fällen den Populationsmittelwert μ enthalten.

Student	Schlafdauer [in Stunden]
1	7
2	9
3	7
4	7
5	9
6	8
7	6
8	7
9	8
10	7
11	7
12	9
13	8
14	9
15	6
16	8
17	7
18	7
19	6
20	8

Tab. 4: Umfragewerte von Lena und Hasso. Sie bilden das approximative Modell der Population.

Fall 1:

Hasso ruft bei der Schulleitung an und erhält den Tipp $b = 0{,}4$. Anhand des Tipps ist die Lösung das Intervall $7{,}5 \pm 0{,}4 = [7{,}1;\ 7{,}9]$. In 95 % der Fälle liegt der Populationsmittelwert μ im Intervall $[\bar{x} - b; \bar{x} + b]$. Denn: In 95 % der Fälle liegt der Umfragemittelwert $\bar{x}$ im 95 %-Schwankungsbereich und das Intervall $[\bar{x} - b, \bar{x} + b]$ enthält μ in diesen Fällen (siehe Abbildung 1).

$$I = [\bar{x} - b; \bar{x} + b] = [7{,}1;\ 7{,}9]$$

bezeichnen wir deshalb als genaues 95 %-Konfidenzintervall für den Populationsmittelwert μ.

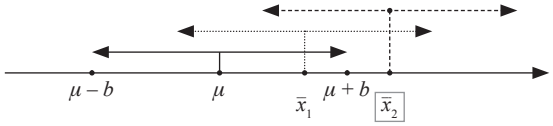


Abb. 1: Das als durchgezogene Linie dargestellte Intervall ist der 95 %-Schwankungsbereich $[\mu - b; \mu + b]$ der Mittelwerte der Stichproben. Die gepunktete Linie stellt das 95 %-Konfidenzintervall $[\bar{x}_1 - b; \bar{x}_1 + b]$ dar und enthält μ, da $\bar{x}_1$ im 95 %-Schwankungsbereich von μ liegt. Die gestrichelte Linie ist das 95 %-Konfidenzintervall $[\bar{x}_2 - b; \bar{x}_2 + b]$; es enthält nicht μ, da $\bar{x}_2$ nicht im 95 %-Schwankungsbereich von μ liegt.

Fall 2:

Lena und Hasso erhalten keinen Tipp von der Schulleitung. Da das Modell der Population nicht bekannt ist, ist auch b nicht bekannt. Deshalb verwenden sie die Umfragewerte, d. h. ihre Stichprobe (Tabelle 4) als approximatives Modell der Population an Stelle des unbekannten exakten Modells der Population (Tabelle 2). Sie ziehen per Simulation wiederholt Stichproben *(Bootstrap-Stichproben)* vom Umfang $n = 20$ aus der vorhandenen Stichprobe um b zu approximieren, d. h.

$$b = [q(0{,}975) - q(0{,}025)]/2$$

wird geschätzt durch

$$\hat{b} = [\hat{q}(0{,}975) - \hat{q}(0{,}025)]/2\,.$$

Dabei ist $\hat{q}(0{,}975)$ bzw. $\hat{q}(0{,}025)$ das 97,5 % bzw. 2,5 % Quantil der Verteilung der Mittelwerte, die anhand des approximativen Modells der Population simuliert wurden (siehe Tabelle 5). Die beiden unbekannten Quantile sowie der Populationsmittelwert werden somit ersetzt durch die entsprechenden Quantile der Bootstrap-Verteilung und den Mittelwert der vorliegenden Daten.

In unserem Fall ist

$$\hat{b} = \frac{7{,}9 - 7{,}15}{2} = 0{,}375$$

und wir erhalten das approximative 95 % Konfidenzintervall $\tilde{I} = [\bar{x} - \hat{b}; \bar{x} + \hat{b}] = [7{,}125; 7{,}875]$ für den Populationsmittelwert. Alternativ hätte man – unter Verzicht auf die Annahme der Symmetrieannahme – auch die entsprechenden Perzentile der Bootstrap-Verteilung als approximatives Konfidenzintervall nehmen können, d. h.

$$\tilde{I} = [\hat{q}(0{,}025); \hat{q}(0{,}975)] = [7{,}15; 7{,}9] .$$

In Abgrenzung zu weiteren im Zusammenhang mit dem Bootstrap stehenden Varianten von Intervallen wird in der Fachliteratur dieses Intervall als Bootstrap-Perzentilintervall *(bootstrap percentile interval)* bezeichnet (Engel 2010).

	A	B	C	D	E	F	G	H	I
1	Daten			Resampling			Wiederholte Mittelwerte		
2									
3	Nummer	Schlafdauer		Nummer	Schlafdauer		7,35		
4	1	7		4	7		7,95		
5	2	9		10	7		7,45		
6	3	7		9	8		7,75		2,5% Quantil
7	4	7		4	7		7,45		7,15
8	5	9		5	9		7,6		
9	6	8		4	7		7,55		
10	7	6		10	7		7,15		
11	8	7		9	8		7,35		97,5% Quantil
12	9	8		7	6		7,35		7,9
13	10	7		10	7		7,65		
14	11	7		9	8		7,15		
15	12	9		10	7		7,4		
16	13	8		6	8		7,3		
17	14	9		3	7		7,75		
18	15	6		6	8		7,5		
19	16	8		4	7		7,7		
20	17	7		8	7		7,55		
21	18	7		3	7		7,4		
22	19	8		10	7		7,5		
23	20	6		6	8		7,4		
24							7,55		
25	Mittelwert	7,5		Mittelwert	7,35		7,25		
26							7,75		
27							7,6		
28							7,65		

Tab. 5: Als approximatives Modell der Population werden die Daten (Spalte B) verwendet. Daraus wird eine neue Stichprobe simuliert (Spalte E) und dessen Mittelwert berechnet (Zelle E25). Das wird 1000 Mal wiederholt und die Mittelwerte werden in Spalte G gespeichert.

4 Bootstrap-Konfidenzintervalle bei asymmetrischen Stichprobenverteilungen

Lena und Hasso untersuchen den Zusammenhang zwischen Motivation und Erfolg für das Studium an der Fachhochschule für Architektur. In einer Umfrage haben sie 12 Mitstudenten nach ihrer Motivation und ihrem Erfolg befragt. Die Motivationswerte können Werte zwischen 0 und 5 annehmen, und die Erfolgswerte haben Werte zwischen 0 und 10. Die Ergebnisse sind in Tabelle 6 zusammengefasst.

Der Pearsonsche Korrelationkoeffizient zwischen Motivation und Erfolg in der Stichprobe beträgt $r = 0{,}69$. Nun wollen Lena und Hasso ein 95 %-Konfidenzintervall für die Korrelation zwischen Motivation und Erfolg in der Population ermitteln.

Im jetzt vorliegenden Fall ist die Stichprobenverteilung asymmetrisch. Somit ist auch die Symmetrieannahme für das Konfidenzintervall nicht mehr haltbar. Im Fall des Pearsonschen Korrelationskoeffizienten ϱ erhält man einen 95 % Schwankungsbereich der Art $[\varrho - b_1 ; \varrho + b_2]$, der nicht um ϱ zentriert ist (siehe Abbildung 2). Dennoch kann man zunächst ganz analog zum vorangegangenen Beispiel vorgehen und Bootstrap-Stichproben (d. h. neue Stichproben aus der vorhandenen Stichprobe) ziehen. Ganz analog wie beim Konfidenzintervall für den Mittelwert erhält man das Bootstrap-Perzentilintervall durch die Perzentile der Bootstrap-Verteilung: $[\hat{q}(0{,}025); \hat{q}(0{,}975)] = [r - \hat{b}_1 ; r + \hat{b}_2]$, wobei $\hat{b}_1 = r - \hat{q}(0{,}025)$; $\hat{b}_2 = \hat{q}(0{,}975) - r$ ist.

Um eine Zufallsstichprobe mit 12 Umfragewerten zu erzeugen, gehen Lena und Hasso daher wie folgt vor: Sie erzeugen eine Zufallszahl zwischen 1 und 12. Sie entspricht dem Studenten aus dem approximativen Modell der Population (die Daten), der befragt wird und schreiben seine Motivation und seinen Erfolg auf. Das wird 12 Mal wiederholt und resultiert somit in 12 gepaarten Umfragewerten, die eine simulierte Zufallsstichprobe bilden (Spalten F und G in Tabelle 7), aus der sich ein Korrelationskoeffizient berechnen lässt (Zelle G17 in Tabelle 7). Dieser Vorgang wird sehr oft (z. B. 1000 mal) wiederholt, wodurch man 1000 anhand der Daten simulierte Korrelationskoeffizienten erhält (Spalte I in Tabelle 7). Im vorliegenden Fall ergibt eine Simulation mit Excel $\hat{q}(0{,}025) = 0{,}28$ und $\hat{q}(0{,}975) = 0{,}92$ und somit das Intervall [0,28; 0,92] als Bootstrap-Perzentil-Intervall

Student	Motivation	Erfolg
1	2	5
2	4	7
3	3	7
4	5	8
5	2	5
6	4	9
7	4	8
8	3	4
9	1	3
10	3	5
11	3	8
12	5	6

Tab. 6: Umfragewerte von Lena und Hasso. Sie bilden das approximative Modell der Population.

	A	B	C	D	E	F	G	H	I	J	K
1	Daten				Resampling				Wiederholte Korrelationen		
2											
3	Student	Motivation	Erfolg		Nummer	Motivation	Erfolg				
4	1	2	5		6	4	9		0.44		2.5% Quantil
5	2	4	7		12	5	6		0.7		0.28
6	3	3	7		4	5	8		0.66		
7	4	5	8		7	4	8		0.68		
8	5	2	5		7	4	8		0.55		97.5% Quantil
9	6	4	9		10	3	5		0.86		0.92
10	7	4	8		10	3	5		0.67		
11	8	3	4		11	3	8		0.62		
12	9	1	3		12	5	6		0.69		
13	10	3	5		4	5	8		0.77		
14	11	3	8		2	4	7		0.68		
15	12	5	6		1	2	5		0.35		
16									0.61		
17		Korrelation	0.69			Korrelation	0.44		0.8		
18									0.53		
19									0.77		

Tab. 7: Als approximatives Modell der Population werden die gepaarten Daten (Spalten B und C) verwendet. Daraus wird eine neue Stichprobe (Spalten F und G) und Korrelation (Zelle G17) simuliert. Das wird 1000 Mal wiederholt und die Korrelationen werden in Spalte I gespeichert.

für den Korrelationskoeffizienten zwischen Motivation und Studienerfolg.

Allerdings ist im Fall asymmetrisch verteilter Schätzer das Bootstrap-Perzentilintervall oft verzerrt, was zu falschen Überdeckungswahrscheinlichkeiten führt (siehe Efron and Tibshirani, S. 178 und Abbildung 2). Eine plausible Alternative im asymmetrischen Fall mit weniger Verzerrung ist das sogenannte klassische Bootstrap-Intervall *(basic bootstrap interval)*. Ausgangspunkt für das klassische Bootstrap-Intervall ist die Differenz zwischen Schätzwert r und Populationsparameter ϱ. Angenommen wir hätten kritische Werte c_1 und c_2, so dass gilt: $P(c_1 \le r - \varrho \le c_2) = 95\,\%$. Wegen

$$\begin{aligned} 95\,\% &= P(q(0{,}025) \le r \le q(0{,}975)) \\ &= P(q(0{,}025) - \varrho \le r - \varrho \le q(0{,}975) - \varrho)) \end{aligned}$$

sind die um ϱ verminderten Perzentile der Stichprobenverteilung von r die exakten Werte für b_1 und b_2. Einfaches Umstellen ergibt $P(r - c_2 \le \varrho \le r - c_1) = 95\,\%$, d. h. $[r - c_2 ; r - c_1]$ ist ein 95 % Konfidenzintervall für ϱ. Da die Verteilung von r aber nicht bekannt ist, ersetzen wir sie durch die Bootstrap-Verteilung, und die entsprechenden Perzentile werden – da ϱ nicht bekannt ist – um r vermindert, d. h. c_1 und c_2 werden ersetzt durch

$$\hat{c}_1 = \hat{q}(0{,}025) - r,\ \hat{c}_2 = \hat{q}(0{,}975) - r.$$

Dies führt schließlich zu

$$\begin{aligned} 95\,\% &= P(r - c_2 \le \varrho \le r - c_1) \\ &\approx P(2r - \hat{q}(0{,}975) \le \varrho \le 2r - \hat{q}(0{,}0025)) \\ &= P(r - \hat{c}_2 \le \varrho \le r - \hat{c}_1). \end{aligned}$$

Demnach ist das approximative 95 %-Konfidenzintervall $\tilde{I} = [r - \hat{c}_2 ; r - \hat{c}_1]$.

Im obigen Beispiel mit den Perzentilen der Bootstrapverteilung von 0,28 bzw. 0,92 ergibt sich für den Korrelationskoeffizienten zwischen Motivation und Studienerfolg [0,46; 1] als klassisches Bootstrap-Konfidenzintervall, wenn man noch berücksichtigt, dass Korrelationskoeffizienten nie größer als 1 werden können.

Visuell lassen sich die Überdeckungswahrscheinlichkeiten folgendermaßen veranschaulichen:

Ist ϱ der Populationskorrelationskoeffizient, dann ist der 95 % Schwankungsbereich $[\varrho - b_1, \varrho + b_2]$, wobei $b_1 = \varrho - q(0{,}025)$, $b_2 = q(0{,}975) - \varrho$.

Ersetzen wir ϱ durch r und die Perzentile durch die entsprechenden Perzentile der Bootstrap-Verteilung, so erhalten wir das Bootstrap-Perzentilintervall, d. h. $\hat{b}_1 = r - \hat{q}(0{,}025)$, $\hat{b}_2 = \hat{q}(0{,}975) - r$. Wie oben erläutert, ist das Intervall $[r - \hat{b}_1, r + \hat{b}_2]$ jedoch verzerrt, d. h. es hat nicht die korrekten Überdeckungswahrscheinlichkeiten. Ein Vergleich mit dem oben hergeleiteten klassischen Bootstrap-Intervall ergibt $\hat{b}_1 = -\hat{c}_1$, $\hat{b}_2 = \hat{c}_2$. Somit ist $[r - \hat{b}_2, r + \hat{b}_1]$ das klassische Bootstrap-Intervall.

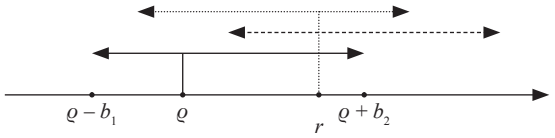


Abb. 2: Das durchgezogene Intervall ist der 95 % Schwankungsbereich $[\varrho - b_1 ; \varrho + b_2]$ der asymmetrisch verteilten Kennzahl ϱ. Eine Stichprobe liefert den Schätzwert r, welcher im 95 % Schwankungsbereich $[\varrho - b_1 ; \varrho + b_2]$ liegt. Das gestrichelte 95 % Bootstrap-Perzentilintervall $[r - b_1 ; r + b_2]$ enthält ϱ nicht, obwohl r im 95 % Schwankungsbereich $[\varrho - b_1 ; \varrho + b_2]$ liegt. Das gepunktete 95 % klassische Bootstrap-Intervall hingegen enthält ϱ.

5 Konvergenz der Methode

Bei der Bestimmung des 95 %-Schwankungsbereiches wurden die Perzentile durch Simulieren von 1000 Zufallsstichproben ermittelt. Dadurch ist eigentlich das hier eingeführte exakte 95 %-Konfidenzintervall – welches anhand der halben Breite $b = q(0{,}975) - \mu$ bestimmt wird – nur näherungsweise exakt. Jedoch hängt die Genauigkeit von b nur von dem Rechenaufwand ab, der durch Erhöhung der Anzahl der Simulationen beliebig erhöht werden kann. Beim Bestimmen des approximativen 95 %-Konfidenzintervalls erscheint eine weitere Fehlerquelle:

Das exakte Modell der Population wird angenähert durch das approximative Modell der Population, also

die Daten. Die Qualität dieser Approximation hängt davon ab, wie weit die Daten repräsentativ für die Grundgesamtheit sind.

Die Daten bilden eine approximative Abbildung der Population. Mit zunehmendem Stichprobenumfang streben die relativen Häufigkeiten der Umfragewerte in der Zufallsstichprobe gegen die relativen Häufigkeiten der Population.

Dadurch strebt $\hat{b} = \hat{q}(0{,}975) - \bar{x}$ gegen $b = q(0{,}975) - \mu$ und somit strebt die Überdeckungshäufigkeit des approximativen 95 %-Konfidenzintervalls gegen 95 %.

Man kann zeigen, dass die Überdeckungshäufigkeit des approximativen $(100 - \alpha)$%-Konfidenzintervalls $1 - \alpha + C/\sqrt{n}$ beträgt, wobei C eine verteilungsabhängige Konstante ist und n den Stichprobenumfang bezeichnet (Hall 1992). Das Bootstrap-Verfahren ist nur so gut, wie es die Repräsentativität der vorliegenden Stichprobe erlaubt, weshalb es bei kleinen Stichproben weniger geeignet ist. Die vorangegangenen Beispiele mit $n = 20$ bzw. $n = 12$ dienten rein illustrativen Zwecken.

6 Verfeinerungen und Erweiterungen

Neben dem Bootstrap-Perzentilintervall und dem klassischen Bootstrap-Intervall gibt es noch viele weitere Varianten (z. B. studentized bootstrap, parametric bootstrap, bias corrected bootstrap, accelerated bootstrap) mit zum Teil deutlich verbesserten Konvergenzeigensschaften. Für Details muss auf die Literatur (z. B. Davison & Hinkely 1997, DiCiccio & Efron1996) verwiesen werden.

Es ist instruktiv, die Breite des Schwankungsbereiches und des Konfidenzintervalls in Abhängigkeit des Stichprobenumfangs und in Abhängigkeit des Konfidenzniveaus (z. B. mit Hilfe von Excel) zu untersuchen und die Ergebnisse dann anschaulich zu deuten. Die Idee, dass das Konfidenzintervall mit zunehmendem Konfidenzniveau breiter und mit zunehmendem Stichprobenumfang feiner wird, kann dann anschaulich interpretiert werden.

Im Rahmen eines statistischen Projektes können Studenten vielfältige Fragestellungen bearbeiten, indem sie zuerst eine Datenerhebung durchführen (z. B. durch eine Umfrage), anhand der beschreibenden Statistik verschiedene Kennwerte ausrechnen und anschließend anhand der hier präsentierten Methoden Konfidenzintervalle für die untersuchten Kennwerte bestimmen.

Literatur

Christie, D. (2004): Resampling mit Excel. In: *Stochastik in der Schule* 24(3), S. 22–27.

Cobb, G. (2007): The introductory statistics course: A Ptolemaic curriculum? In: *Technology Innovations in Statistics Education*, 1(1), S. 1–15.

Davison, A. C. & Hinkley, D. V. (1997): Bootstrap Methods and their Applicatins. Cambridge University Press.

DiCiccio, T. & Efron, B. (1996): Statistical Science 11 (3), 189–228.

Efron, B. (1979): Bootstrap methods: another look at the jackknife. In: *Annals of Statistics* 7(1), S. 1–26.

Efron, B. & Tibshirani, R. (1993): An Introduction to the Bootstrap. Chapman & Hall: London.

Engel, J. (2010): On teaching bootstrap confidence intervals. In: C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute.

Engel, J. & Grübel, R. (2008): Bootstrap – oder die Kunst, sich selbst aus dem Sumpf zu ziehen. In: *Mathematische Semesterberichte*, 55, S. 113–130.

Erickson, T. (2002): Fifty Fathoms. Eeps-media: Oakland

Hall, P. (1992): The Bootstrap and Edgworth Expansion. Springer-Verlag: London, 1992.

Hesterberg, T. (2014): Bootstraping for learning statistics. In: K. Makar; B. de Sousa & R. Gould (Eds.). *Proceedings of the Ninth International Conference on Teaching Statistics*, Voorburg, The Netherlands: International Statistical Institute.

Johnson, R. W. (2001): An Introduction to the Bootstrap. In: *Teaching Statistics* 23(2), S. 49–54.

Pfannkuch, M. & Budgett, S. (2014): Constructing inferential concepts through bootstrap and randomization-test simulations: a case study. In: K. Makar; B. de Sousa & R. Gould (Eds.). *Proceedings of the Ninth International Conference on Teaching Statistics*, Voorburg, The Netherlands: International Statistical Institute.

Watson, J. (2013): Resampling with Tinkerplots. In: *Teaching Statistics* 35(1), S. 32–36.

Anschrift der Verfasser

Andreas Kaufmann
Institut für Allgemeinbildende Fächer
Fachhochschule Westschweiz
CH-1040 Echallens
andreas.kaufmannn@telkel.ch

Joachim Engel
Institut für Mathematik und Informatik
Pädagogische Hochschule Ludwigsburg
Reuteallee 36
D-71634 Ludwigsburg
engel@ph-ludwigsburg.de

Der p-Wert: Standardisierte Zufallsvariable, Überschreitungswahrscheinlichkeit oder Grenzniveau des Ablehnens?

FRANK MAROHN, WÜRZBURG

Zusammenfassung: *In der vorliegenden Arbeit soll einmal mehr auf den p-Wert eingegangen werden. Dieser Wert, der von statistischen Software-Paketen, sprich vom Computer, berechnet wird und auf den sich die statistische Anwenderwelt „stürzt", dient als Entscheidungsgrundlage beim Testen von Hypothesen. Gerade weil der Computer in den Schulen Einzug gehalten hat, ist es für den Lehrenden wichtig zu wissen, was der p-Wert (nicht) ist.*

1 Einleitung

Ein statistischer Test ist eine Entscheidungsregel, die festlegt, ob man sich aufgrund der vorliegenden Daten für die Nullhypothese H_0 oder sich gegen H_0 und somit für die Alternative H_1 entscheidet.

Üblicherweise gibt man sich nach der sogenannten Neyman-Pearson-Methode ein Signifikanzniveau (Testniveau) α vor, das die Wahrscheinlichkeit für einen Fehler 1. Art (fälschliche Ablehnung von H_0) kontrolliert. Im Gegensatz zur Vorgehensweise, einen Höchstwert α für die Wahrscheinlichkeit des Fehlers 1. Art festzulegen und daraufhin den kritischen Bereich zu wählen, ist es insbesondere bei der Verwendung von statistischer Software gängige Praxis, aus einer Stichprobe einen sogenannten p-Wert auszurechnen und die Signifikanz des erhaltenen Resultates anhand dieses Wertes zu beurteilen. Ein kleiner p-Wert (üblicherweise ≤ 0.05) führt zu einer Ablehnung von H_0.

Was ist der p-Wert? Ist diese Zahl eine Realisierung einer Zufallsvariablen? Ist sie eine Wahrscheinlichkeit? Oder ein Signifikanzniveau? Auf diese Fragen soll in diesem Artikel näher eingegangen werden. In der Literatur finden sich folgende drei Definitionen:

(1) *Der p-Wert als Zufallsvariable:* „Der p-Wert ist keine Wahrscheinlichkeit" (Stahel 2008, S. 209). Der p-Wert ist eine Zufallsvariable bzw. der konkrete p-Wert (also die Zahl, die vom Computer berechnet wird) ist eine Realisierung dieser Zufallsvariablen (Stahel 2008, Abschnitt 8.3; Falk et al. 2004, Abschnitt 2.3).

(2) *Der p-Wert als Wahrscheinlichkeit:* Der p-Wert ist die Wahrscheinlichkeit, unter H_0 den beobachteten Prüfgrößenwert oder einen in Richtung der Alternative extremeren Wert zu erhalten (Freund & Perles 1996; Stahel 2008, Abschnitt 8.7; Rudolf & Kuhlisch 2008, Abschnitt 5.3.1).

(3) *Der p-Wert als Grenzniveau:* Der p-Wert zu einer Beobachtung ist das kleinste Signifikanzniveau, zu dem der Test H_0 verwirft (Falk et al. 2014, Abschnitt 2.7.1; Henze 2013, Abschnitt 29.4; Krengel 2005, Abschnitt 6.10; Stahel 2008, Abschnitt 8.3).

Auf diese drei Definitionen des p-Wertes soll im Folgenden genauer eingegangen werden. In Abschnitt 2 wird zunächst der Fall einer stetig verteilten Prüfgröße betrachtet. Als Referenz dient der Einstichproben-Gauß-Test. Der Fall einer diskret verteilten Prüfgröße wird in Abschnitt 3 behandelt. Als Referenz dient der Binomialtest. Bezüglich des Gauß-Tests und des Binomialtests sei auf Georgii 2009 und Henze 2013 verwiesen. Aspekte rund um den p-Wert werden in Abschnitt 4 diskutiert.

2 Der p-Wert im Stetigen

Wir betrachten den Fall einer stetig verteilten Prüfgröße. Man nennt eine Zufallsvariable X stetig verteilt, falls eine Funktion $f \geq 0$ existiert (die sogenannte Wahrscheinlichkeitsdichte) mit $P(a \leq X \leq b) = \int_a^b f(x)\,dx$, $a < b$. Die Verteilungsfunktion $F(x) = \int_{-\infty}^{x} f(t)\,dt$ ist dann stetig. Exemplarisch betrachten wir eine normalverteilte Prüfgröße.

2.1 Der p-Wert als Zufallsvariable

In diesem Abschnitt werden wir sehen, dass der p-Wert eine Art „vollstandardisierte" Prüfgröße ist, die auf $(0,1)$ gleichverteilt ist, vorausgesetzt, die Verteilungsfunktion der Prüfgröße ist stetig.

Eine Zufallsvariable U heißt auf $(0,1)$ *gleichverteilt* (uniformly distributed), falls U die Dichte $f(x) = 1$, $x \in (0,1)$ und $f(x) = 0$, $x \notin (0,1)$, besitzt. Für ein Intervall $(a,b) \subset (0,1)$ gilt somit

$$P(a \leq U \leq b) = \int_a^b f(x)\,dx = b - a.$$

Speziell gilt $P(U \leq 0.05) = 0.05$. Die Wahrscheinlichkeit, dass eine auf $(0,1)$ gleichverteilte Zufallsvariable einen Wert im Intervall $(0, 0.05)$ anneh-

men wird, beträgt also 0.05. Die folgende allgemeine Aussage ist für den p-Wert wichtig.

Transformation in die Gleichverteilung: Besitzt eine Zufallsvariable X eine stetige Verteilungsfunktion F, so ist $F(X)$ auf (0,1) gleichverteilt.

Der Beweis dieser Aussage, bei der man nicht auf die Stetigkeit von F verzichten kann, ist einfach, falls F streng monton wachsend ist. Denn in diesem Fall existiert die Umkehrfunktion F^{-1}. Im allgemeinen Fall hat man die sogenannte Quantiltransformation zu betrachten, die bei der Erzeugung von (Pseudo)-Zufallszahlen eine grundlegende Rolle spielt. Für Details und die Beweise sei auf Georgii 2009, Kapitel 1 und auf Henze 2013, Abschnitt 31.14, verwiesen.

Im Folgenden wollen wir den Einstichproben-Gauß-Test betrachten, gehen also von einem Normalverteilungsmodell $N(\mu,\sigma^2)$ mit unbekanntem Mittelwert $\mu\in\mathbb{R}$ und *bekannter* Varianz $\sigma^2>0$ aus. Betrachtet wird zunächst das linksseitige Testproblem

$$\text{(L)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu<\mu_0.$$

Wir können genauso gut $H_0:\mu\geq\mu_0$ schreiben. Bei der Fehlerwahrscheinlichkeit 1. Art kommt es nur auf den „Randpunkt" μ_0 an, der die Nullhypothese von der Alternative trennt. Einseitigkeit bezieht sich auf die Alternative.

Basierend auf einer unabhängig und identisch verteilten Stichprobe $X_1,\ldots,X_n$, lautet die Prüfgröße

$$Z=\frac{\bar{X}_n-\mu_0}{\sigma/\sqrt{n}}.$$

Dabei bezeichnet wie üblich $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ das Stichprobenmittel. Unter H_0 besitzt Z eine $N(0,1)$-Verteilung. Bezeichnet

$$\Phi(z)=\int_{-\infty}^{z}\varphi(x)\,dx \text{ mit } \varphi(x):=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

die Verteilungsfunktion der $N(0,1)$-Verteilung.

Der p-Wert zum Testproblem (L) ist gegeben durch die Zufallsvariable

$$p_L(Z):=\Phi(Z)=\int_{-\infty}^{Z}\varphi(x)\,dx$$

(Verknüpfung von Φ und Z), welche auf $(0,1)$ gleichverteilt ist. Speziell gilt $P_{\mu_0}(p_L(Z)\leq 0.05)=0.05$. Die Indizierung von P mit μ_0 bedeutet, dass bei der Berechnung der Wahrscheinlichkeit der Parameter μ_0 unterstellt wird.

Ist z der konkret beobachtete Prüfgrößenwert (Realisierung von Z), basierend auf Daten $x_1,\ldots,x_n$ (Realisierungen von $X_1,\ldots,X_n$), so ist $p_L(z)=\Phi(z)$ eine Realisierung der Zufallsvariablen $\Phi(Z)$. Die Zahl $p_L(z)$, die vom Computer berechnet wird, ist der p-Wert zur Beobachtung z. Da die Werte der Verteilungsfunktion Φ tabelliert sind, lässt sich im Fall des Gauß-Tests der p-Wert zur Beobachtung z aus den entsprechenden Tabellenwerken für Φ ablesen.

Betrachtet man das rechtsseitige Testproblem

$$\text{(R)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu>\mu_0,$$

so ist

$$p_R(Z):=1-\Phi(Z)=\int_{Z}^{\infty}\varphi(x)\,dx$$

der p-Wert und $p_R(z)$ ist der p-Wert zur Beobachtung z. Beachte: Mit $\Phi(Z)$ ist auch $1-\Phi(Z)$ auf $(0,1)$ gleichverteilt, so dass wieder $P_{\mu_0}(p_R(Z)\leq 0.05)=0.05$ gilt.

Im Testproblem (L) sprechen kleine z-Werte und im Testproblem (R) sprechen große z-Werte für die Alternative. Zur Beurteilung, ob ein kleiner oder großer Prüfgrößenwert beobachtet worden ist, braucht es eine Bezugsgröße, da die Prüfgröße Z im Prinzip beliebig kleine oder große Werte annehmen kann. Wird ein Signifikanzniveau α vorgegeben, so bewertet man nach der Neyman-Pearson-Methode den Wert z durch den Vergleich mit einem kritischen Wert, sprich Quantil (wir kommen in Abschnitt 2.3 darauf zurück).

Eine andere Möglichkeit (und dies ist die p-Wert-Methode) besteht darin, die Prüfgröße so zu transformieren, dass sie beschränkt ist. Die Schranken dienen dann als Bezugsgrößen. Durch die Transformation $Z\mapsto\Phi(Z)$ bzw. $Z\mapsto 1-\Phi(Z)$ der Prüfgröße Z auf ihren p-Wert lassen sich auffällig kleine bzw. große Realisierungen von Z *unmittelbar* erkennen. Übliche Konvention: p-Werte kleiner oder gleich 0.05 sprechen gegen die Nullhypothese. Fazit: Der p-Wert ist eine Art „vollstandardisierte" Prüfgröße, die unter H_0 eine uniforme Verteilung besitzt.

Beim beidseitigen (oder zweiseitigen) Testproblem

$$\text{(B)}\quad H_0:\mu=\mu_0 \quad\text{vs}\quad H_1:\mu\neq\mu_0$$

ist der p-Wert zur Beobachtung z definiert durch

$$\begin{aligned} p_B(z) &:= 2\cdot\min\{(p_L(z),p_R(z)\} \\ &= 2\cdot\min\{\Phi(z),1-\Phi(z)\} \end{aligned} \qquad (1)$$

Die Zufallsvariable $p_B(Z)$ ist ebenfalls auf $(0,1)$ gleichverteilt. Aufgrund der Symmetrie der Dichte φ

können wir für den zweiseitigen p-Wert auch schreiben $p_B(z) = 2 \cdot (1 - \Phi(|z|))$.

Bemerkung 1: Auch bei nichtsymmetrischen Verteilungen (z. B. Chi-Quadrat-Verteilung) ist der beidseitige p-Wert definiert als das Doppelte des Minimums der beiden einseitigen p-Werte.

2.2 Der p-Wert als Wahrscheinlichkeit

In diesem Abschnitt wollen wir eine (die einzige?) Möglichkeit kennenlernen, den p-Wert mit dem Begriff einer Wahrscheinlichkeit in Zusammenhang zu bringen. Betrachten wir dazu das Testproblem (R). Häufig liest man dann die folgende (etwas laxe) Definition: Der p-Wert zur Beobachtung z ist die Wahrscheinlichkeit, unter der Nullhypothese einen Prüfgrößenwert zu beobachten, der größer oder gleich z ist. Formal wird in der Literatur für diese *Überschreitungswahrscheinlichkeit* $P_{\mu_0}(Z \geq z)$ geschrieben. Diese Definition des p-Wertes zur Beobachtung z ist gleichbedeutend mit der in Abschnitt 2.1, da Z unter H_0 die Verteilungsfunktion Φ besitzt, d.h. $\Phi(z) = P_{\mu_0}(Z \leq z)$, $z \in \mathbb{R}$:

$$P_{\mu_0}(Z \geq z) = 1 - \Phi(z) = p_{\mathrm{R}}(z).$$

An dieser Stelle sei zur Notation Folgendes gesagt:

1. Wir geben der ebenfalls in der Literatur üblichen Bezeichnungsweise $P(Z \geq z|\mu_0)$ gegenüber $P_{\mu_0}(Z \geq z)$ nicht den Vorzug, da diese Notation eine bedingte Wahrscheinlichkeit suggerieren würde. Hypothesen, also Modelle, haben keine Wahrscheinlichkeiten, sie legen Wahrscheinlichkeiten fest (siehe Bemerkung 3).

2. Unreflektiert scheint die Schreibweise für den p-Wert als Überschreitungswahrscheinlichkeit $P_{\mu_0}(Z \geq z)$ unproblematisch zu sein. Aber genauer betrachtet ergibt sich hierbei die folgende Schwierigkeit: Die Beobachtung z ist eine Realisierung von Z. Die Daten $x_1, \ldots, x_n$ (Realisierungen von $X_1, \ldots, X_n$), die zum Prüfgrößenwert z führen, liegen vor. Das Zufallsexperiment ist also *bereits durchgeführt* worden (Vergangenheit!). Daher macht der Ausdruck $P_{\mu_0}(Z \geq z)$ wenig Sinn, denn Wahrscheinlichkeitsaussagen können sich nur auf künftige (oder nicht bekannte) Ereignisse beziehen. Wohin die laxe Schreibweise $P_{\mu_0}(Z \geq z)$ führt wird deutlich, wenn wir versuchen, den p-Wert wie in Abschnitt 2.1 als Zufallsvariable zu schreiben: $1 - \Phi(Z) = P_{\mu_0}(Z \geq Z) = 1$. Unproblematisch ist dagegen die Schreibweise

$$p_{\mathrm{R}}(z) = P_{\mu_0}(\tilde{Z} \geq z).$$

Dabei besitzt $\tilde{Z}$ unter P_{μ_0} die gleiche Verteilung wie die Prüfgröße Z und $\tilde{Z}, Z$ sind stochastisch unabhängig. Mit anderen Worten: Wenn man das gleiche Zufallsexperiment noch einmal unabhängig wiederholen würde, so ist der p-Wert die Wahrscheinlichkeit, dass unter der Nullhypothese die Prüfgröße $\tilde{Z}$ einen Wert annehmen wird, der größer oder gleich z ist. Jetzt können wir auch den p-Wert als Zufallsvariable schreiben:

$$p_{\mathrm{R}}(Z) = 1 - \Phi(Z) = P_{\mu_0}\big(\tilde{Z} \geq Z\big). \qquad (2)$$

Bemerkung 2: Man kann den p-Wert sehen als bedingte Wahrscheinlichkeit (Übergangswahrscheinlichkeit) im Rahmen eines zweistufigen Zufallsexperimentes. Es gilt dann

$$P_{\mu_0}\big(\tilde{Z} \geq Z | Z = z\big) = P_{\mu_0}\big(\tilde{Z} \geq z\big) = p_{\mathrm{R}}(z). \qquad (3)$$

Was sich so einfach liest und einleuchtend erscheint, ist alles andere als offensichtlich. Erstens: Es handelt sich hier um ein allgemeines Konzept von bedingten Verteilungen und nicht um den Begriff einer elementaren bedingten Wahrscheinlichkeit, wie man ihn aus der Schule kennt (beachte: $P_{\mu_0}\{Z = z\} = 0$). Zweitens: Die Gültigkeit des ersten Gleichheitszeichens in (3) ist nicht selbstverständlich. Diese „Einsetzungsregel", so intuitiv klar sie auch sein mag, bedarf eines Beweises (siehe z. B. Wengenroth 2008, Satz 6.4).

Halten wir fest: Der rechtsseitige p-Wert kann als Überschreitungswahrscheinlichkeit interpretiert werden, wobei H_0 zugrunde liegt. Eine schlimme Fehlinterpretation ist die Folgende: Ein p-Wert von 0.042 besagt, dass „die Nullhypothese die Wahrscheinlichkeit 0.042 hat". Eine solche Aussage ist unsinnig. Modelle (und Parameterwerte) selber haben keine Wahrscheinlichkeiten, sie legen Wahrscheinlichkeiten (für Daten und Teststatistiken) fest!

Bemerkung 3: In der Bayes'schen Statistik haben auch Hypothesen bzw. Parameter eine Verteilung, aber da gibt es keine p-Werte. Beachte: Der p-Wert unterstellt die Gültigkeit der Nullhypothese und ist somit nicht als a-posteriori-Wahrscheinlichkeit zu interpretieren.

Entsprechendes gilt für die links- bzw. beidseitige Testsituation. Der p-Wert lässt sich dann interpretieren als Unterschreitungswahrscheinlichkeit $p_{\mathrm{L}}(z) = P_{\mu_0}\big(\tilde{Z} \leq z\big)$ bzw. als Überschreitungswahrscheinlichkeit $p_{\mathrm{B}}(z) = P_{\mu_0}\big(|\tilde{Z}| \geq |z|\big)$.

2.3 Der p-Wert als Grenzniveau

Nach der Neyman-Pearson-Methode wird durch die Vorgabe eines Signifikanzniveaus α die Wahrschein-

lichkeit eines Fehlers 1. Art kontrolliert. Bleiben wir beim Testproblem (R). Der kritische Wert ist das $(1-\alpha)$-Quantil der $N(0,1)$-Verteilung, kurz $z_{1-\alpha}$. Diese Zahl ist per Definition die Lösung der Gleichung $\Phi(x) = 1-\alpha$. Der kritische Bereich (= Ablehnungsbereich) ist somit das Intervall $K_R(\alpha) := [z_{1-\alpha}, \infty)$.

Die Wahrscheinlichkeit eines Fehlers 1. Art ist (höchstens) α:

$$P_{\mu_0}(Z \geq z_{1-\alpha}) = 1 - \Phi(z_{1-\alpha}) = \alpha$$

(für $\mu < \mu_0$ gilt $P_\mu(Z \geq z_{1-\alpha}) < \alpha$).

Die p-Wert-Methode fragt nach dem kleinsten kritischen Bereich, der bei Vorliegen des beobachteten Prüfgrößenwertes z zu einer Ablehnung von H_0 führen würde. Alle sinnvollen kritischen Bereiche sind von der Form $[c, \infty)$, da große Prüfgrößenwerte für die Alternative sprechen. Der kleinste kritische Bereich, der z enthält, ist $[z, \infty)$. Das „Signifikanzniveau" $\alpha^*(z)$, das zu diesem kritischen Bereich gehört, ist der p-Wert. Wir können also wieder schreiben

$$\alpha^*(z) = P_{\mu_0}\left(\tilde{Z} \geq z\right) = p_{\mathrm{R}}(z).$$

Der p-Wert ist ein „Grenzniveau". Es ist die *größte untere Schranke* für das Signifikanzniveau, zu dem H_0 bei Vorliegen des Prüfgrößenwertes z (also im Nachhinein betrachtet) hätte abgelehnt werden können. Es gilt

$$p_R(z) \begin{cases} < \alpha, & z > z_{1-\alpha} \\ = \alpha, & z = z_{1-\alpha} \\ > \alpha, & z < z_{1-\alpha} \end{cases}$$

Daraus folgt: Der p-Wert ist kleiner oder gleich α genau dann, wenn der Prüfgrößenwert z im kritischen Bereich liegt:

$$p_R(z) \leq \alpha \Leftrightarrow z \in K_R(\alpha).$$

Für die Testentscheidung bedeutet dies folgendes: Nach der Neyman-Pearson-Methode wird die Nullhypothese H_0 zum Signifikanzniveau α abgelehnt, wenn der Prüfgrößenwert in den kritischen Bereich fällt; dies ist gleichbedeutend damit, dass der p-Wert kleiner oder gleich α ist.

Schauen wir uns noch kurz das beidseitige Testproblem (B) an. In diesem Fall ist der kritische Bereich zum Signifikanzniveau α gegeben durch

$$K_B(\alpha) := (-\infty, -z_{1-\alpha/2}] \cup [z_{1-\alpha/2}, \infty).$$

Alle kritischen Bereiche sind von Form $(-\infty, -c] \cup [c, \infty)$, $c > 0$. Der kleinste kritische Bereich, der die Beobachtung z enthält, ist gegeben durch $(-\infty, -|z|] \cup [|z|, \infty)$. Das „Signifikanzniveau" dieses kritischen Bereiches ist dann wieder der p-Wert:

$$\begin{aligned} p_B(z) &= 2 \cdot \Phi(-|z|) \\ &= 2 \cdot (1 - \Phi(|z|) \\ &= 2 \cdot \min\{\Phi(z), 1 - \Phi(z)\} \end{aligned}$$

Wegen

$$p_B(z) \begin{cases} < \alpha, & |z| > z_{1-\alpha/2} \\ = \alpha, & |z| = z_{1-\alpha/2} \\ > \alpha, & |z| < z_{1-\alpha/2} \end{cases}$$

führt auch hier die Neyman-Pearson-Methode und die p-Wert-Methode zur gleichen Testentscheidung: Ablehnung von H_0, falls

$$p_B(z) \leq \alpha \Leftrightarrow z \in K_B(\alpha)$$

Aufgrund der Eigenschaft eines „Grenzniveaus" in Abhängigkeit von der Beobachtung wird der p-Wert in der Literatur auch als *tatsächliches*, *exaktes*, *beobachtetes* oder *empirisches* Signifikanzniveau bezeichnet.

Aber Achtung! Der p-Wert ist nicht als ein Signifikanzniveau zu interpretieren (deshalb ist es besser, beim p-Wert von einer größten untere Schranke für das Signifikanzniveau zu sprechen). Das Signifikanzniveau α charakterisiert einen statistischen Test in dem Sinne, dass *bei Unterstellung der Gültigkeit von* H_0 die Wahrscheinlichkeit für eine Ablehnung von H_0 (Fehler 1. Art) höchstens α ist. D. h., in vielen Testdurchführungen wird es unter H_0 in etwa $\alpha \cdot 100\%$ der Fälle zu einer (fälschlichen) Ablehnung von H_0 kommen.

Der p-Wert entzieht sich einer solchen *frequentistischen* Interpretation, da er von den Daten abhängt. Aus diesem Grunde kann der p-Wert auch nicht als eine Wahrscheinlichkeit für den Fehler 1. Art interpretiert werden. Die Aussage „Die Irrtumswahrscheinlichkeit ist gleich 0.042" ist also falsch. Die Irrtumswahrscheinlichkeit charakterisiert einen Test (Nullhypothese, kritischer Bereich) und hat nichts mit Daten zu tun.

3 Der p-Wert im Diskreten

In diesem Abschnitt wollen wir auf den p-Wert eingehen, wenn die Prüfgröße diskret verteilt ist. Dann kann der p-Wert als Zufallsvariable keine (stetige) Gleichverteilung auf $(0,1)$ besitzen. Wir betrachten im Folgenden das Binomialmodell $B_{n,\theta}$, $\theta \in (0,1)$, stellvertretend für diskrete Modelle. Die Binomialverteilung $B_{n,\theta}$ ist ein Wahrscheinlichkeitsmaß auf (der Potenzmenge von) $\{0,\dots,n\}$, festgelegt durch die Einzelwahrscheinlichkeiten

$$B_{n,\theta}(\{j\}) := \binom{n}{j}\theta^j(1-\theta)^{n-j}, \quad j = 0,\dots,n.$$

Dieses Wahrscheinlichkeitsmaß taucht auf als Verteilung der zufälligen Anzahl von Treffern X in einer Bernoulli-Kette der Länge n. Unter der Trefferwahrscheinlichkeit $\theta \in (0,1)$ besitzt X eine $B_{n,\theta}$-Verteilung, d. h.

$$P_\theta(X = j) = B_{n,\theta}(\{j\}), \quad j = 0,\dots,n.$$

Wir betrachten zunächst nur die einseitige Testsituation.

3.1 Der p-Wert als Zufallsvariable

Gegeben sei das linksseitige Testproblem

$$(\mathrm{L}') \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0$$

Eine geeignete Prüfgröße ist X, die unter H_0 B_{n,θ_0}-verteilt ist. Bezeichne

$$F_{\theta_0}(x) := \sum_{j=0}^{[x]} B_{n,\theta_0}(\{j\})$$

die Verteilungsfunktion der B_{n,θ_0}-Verteilung. Dabei ist $[x]$ der ganzzahlige Teil einer reellen Zahl x. Dies ist eine Treppenfunktion mit den Sprungstellen j und Sprunghöhen $B_{n,\theta_0}(\{j\})$, $j = 0,\dots,n$.

Sei k die beobachtete Anzahl von Treffern (Realisierung von X). Dann ist der p-Wert zur Beobachtung k gegeben durch

$$p_{L'}(k) := F_{\theta_0}(k) = \sum_{j=0}^{k} B_{n,\theta_0}(\{j\}),$$

als Zufallsvariable geschrieben

$$p_{L'}(X) := F_{\theta_0}(X) = \sum_{j=0}^{X} B_{n,\theta_0}(\{j\})$$

Diese ist diskret verteilt und kann daher nicht mehr auf $(0,1)$ gleichverteilt sein. Die Verteilungsfunktion von $F_{\theta_0}(X)$ ist eine Treppenfunktion, deren Graph immer unterhalb der Diagonalen liegt. Speziell gilt $P_{\theta_0}\big(p_{L'}(X) \le 0.05\big) \le 0.05$.

Im Fall des rechtsseitigen Testproblems

$$(R') \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

ist

$$p_{\mathrm{R}'}(k) := 1 - F_{\theta_0}(k-1) = \sum_{j=k}^{n} B_{n,\theta_0}(\{j\})$$

der p-Wert zur Beobachtung k.

3.2 Der p-Wert als Wahrscheinlichkeit

Der p-Wert kann wieder als Wahrscheinlichkeit interpretiert werden: Etwa beim Testproblem (R') ist der p-Wert die Überschreitungswahrscheinlichkeit $P_{\theta_0}\big(\tilde{X} \ge k\big)$, also die Summe der Einzelwahrscheinlichkeiten $\sum_{j=k}^{n} P_{\theta_0}\big(\tilde{X} = j\big)$. Dies ist die Wahrscheinlichkeit, unter θ_0 mindestens k Treffer zu beobachten. Dabei bezieht sich die zufällige Trefferanzahl $\tilde{X}$ wieder auf die unabhängige Wiederholung des gleichen Zufallsexperiments (Bernoulli-Kette der Länge n). Formal können wir den p-Wert als Übergangswahrscheinlichkeit, sprich (elementare) bedingte Wahrscheinlichkeit, auffassen:

$$p_{\mathrm{R}'}(k) = P_{\theta_0}\big(\tilde{X} \ge X | X = k\big) = P_{\theta_0}\big(\tilde{X} \ge k\big).$$

3.3 Der p-Wert als Grenzniveau

Auch als Grenzniveau ergibt sich der p-Wert. Bleiben wir beim Testproblem (R'). Alle sinnvollen kritischen Bereiche sind von der Form $\{l,\dots,n\}$, da große Trefferzahlen für die Alternative sprechen.

Ist ein Signifikanzniveu α vorgegeben, so wählt man den kritischen Bereich $K_{\mathrm{R}'}(\alpha) := \{c_\alpha,\dots,n\}$ mit dem kritischen Wert

$$c_\alpha = \min\big\{l \in \{0,\dots,n\} : P_{\theta_0}\big(X \ge l\big) \le \alpha\big\}.$$

Der kleinste kritische Bereich, der die Beobachtung k enthält, ist $\{k,\dots,n\}$. Das „Signifikanzniveau" $\alpha^*(k)$, das zu diesem kritischen Bereich gehört, ist der p-Wert. Wir können also wieder schreiben

$$\alpha^*(k) = P_{\theta_0}\big(\tilde{X} \ge k\big) = p_{\mathrm{R}'}(k).$$

Es ist die größte untere Schranke für das Signifikanzniveau, zu dem H_0 bei Vorliegen des Beobachtungswertes k abgelehnt werden kann:

$$p_{\mathrm{R}'}(k) \begin{cases} \le \alpha, & k \ge c_\alpha \\ > \alpha, & k < c_\alpha \end{cases}$$

Bezüglich der Testentscheidung stimmen p-Wert-Methode und Neyman-Pearson-Methode überein:

$$p_{\mathrm{R}'}(k) \leq \alpha \Leftrightarrow k \in K_{\mathrm{R}'}(\alpha)$$

Gilt dies auch bei zweiseitigen Fragestellungen? Die salomonische Antwort lautet: Es kommt drauf an. Die Antwort hängt davon ab, wie man den p-Wert definiert.

3.4 Festlegung von p-Werten im beidseitigen Testproblem

In der beidseitigen Testsituation

$$(\mathrm{B}') \quad H_0 : \theta = \theta_0 \quad \text{gegen} \quad H_1 : \theta \neq \theta_0$$

wollen wir zunächst den symmetrischen Fall $\theta_0 = 0.5$ behandeln. In diesem Fall gibt es nur eine sinnvolle Festlegung des p-Wertes. Sie ist gegeben durch

$$p_{\mathrm{B}'} = \begin{cases} 2 \cdot F_{\theta_0}(k), & k < n/2 \\ 2 \cdot \big(1 - F_{\theta_0}(k-1)\big), & k > n/2 \\ 1, & k = n/2 \end{cases} \qquad (4)$$

Der zugehörige kritische Bereich ist

$$\begin{aligned} \{0,\ldots,k\} \cup \{n-k,\ldots,n\}, & \quad k < n/2 \\ \{0,\ldots,n-k\} \cup \{k,\ldots,n\}, & \quad k > n/2 \\ \{0,\ldots,n\}, & \quad k = n/2 \end{aligned}$$

Im nichtsymmetrischen Fall $\theta_0 \neq 0.5$ gibt es - anders als im stetigen Fall - verschiedene, aber gleichberechtigte Festlegungen des p-Wertes. Dies kann zu unterschiedlichen Testentscheidungen führen (Abschnitt 3.5) und es kann zu Abweichungen zwischen verschiedenen Statistik-Programmen kommen (Abschnitt 4.1).

Wir erwähnen die folgenden drei Definitionen, die im symmetrischen Fall mit (4) überein stimmen:

1. Version: Man definiert den beidseitigen p-Wert als das doppelte des Minimums vom linksseitigen und rechtsseitigen p-Wert:

$$\begin{aligned} p_{\mathrm{B}'}(k) &:= 2 \cdot \min\{F_{\theta_0}(k), 1 - F_{\theta_0}(k-1), 0.5\} \\ &= 2 \cdot \min\{P_{\theta_0}(\tilde{X} \leq k), P_{\theta_0}(\tilde{X} \geq k), 0.5\} \end{aligned}$$

2. Version: Realisierungen werden als extrem bezeichnet, wenn sie betragsmäßig mehr vom Erwartungswert $E_{\theta_0}(\tilde{X}) = n\theta_0$ abweichen als die Beobachtung k. Der beidseitige p-Wert ist somit

$$\begin{aligned} & p_{\mathrm{B}'}(k) \\ &:= P_{\theta_0}\big(|\tilde{X} - n\theta_0| \geq |k - n\theta_0|\big) \\ &= \begin{cases} P_{\theta_0}(\tilde{X} \leq k) + P_{\theta_0}(\tilde{X} \geq 2n\theta_0 - k), & k < n\theta_0 \\ P_{\theta_0}(\tilde{X} \geq k) + P_{\theta_0}(\tilde{X} \leq 2n\theta_0 - k), & k > n\theta_0 \\ 1, & k = n\theta_0 \end{cases} \end{aligned}$$

3. Version: Bei dieser Festlegung gilt eine Beobachtung als extrem, wenn sie eine kleine Eintrittswahrscheinlichkeit besitzt. Man addiert alle Trefferwahrscheinlichkeiten $P_{\theta_0}\big(\tilde{X} = j\big)$ auf, die kleiner oder gleich $P_{\theta_0}\big(\tilde{X} = k\big)$ sind:

$$\begin{aligned} & p_{\mathrm{B}'}(k) \\ &:= \sum_{\substack{j \in \{0,\ldots,n\} \\ P_{\theta_0}(\tilde{X}=j) \leq P_{\theta_0}(\tilde{X}=k)}} P_{\theta_0}(\tilde{X} = j) \\ &= \begin{cases} P_{\theta_0}(\tilde{X} \leq k) + P_{\theta_0}(\tilde{X} \geq k_1), & k < n\theta_0 \\ P_{\theta_0}(\tilde{X} \geq k) + P_{\theta_0}(\tilde{X} \leq k_2), & k > n\theta_0 \\ 1, & k = n\theta_0 \end{cases} \end{aligned}$$

Dabei sind

$$\begin{aligned} k_1 &= \min\{j > n\theta_0 : P_{\theta_0}(\tilde{X} = j) \leq P_{\theta_0}(\tilde{X} = k)\} \\ k_2 &= \max\{j < n\theta_0 : P_{\theta_0}(\tilde{X} = j) \leq P_{\theta_0}(\tilde{X} = k)\} \end{aligned}$$

Allen drei Versionen ist gemeinsam, dass der p-Wert die Wahrscheinlichkeiten an den Flanken betrachtet. Die Funktion $j \mapsto P_\theta\big(\tilde{X} = j\big)$, $j = 0,\ldots,n$, ist nämlich erst streng monoton steigend, dann streng monoton fallend (für Details siehe Georgii 2009, Lemma 8.8). Die Versionen können, müssen aber nicht zum selben p-Wert führen.

3.5 Vergleich der Testentscheidungen nach NP und der p-Wert-Methode

Nach der Neyman-Pearson-Methode wird H_0 zum Niveau α abgelehnt, falls $k \in K_{\mathrm{B}'}(\alpha) := \{0,\ldots,c_{1,\alpha}\} \cup \{c_{2,\alpha},\ldots,n\}$. Dabei sind die kritischen Werte definiert durch $c_{1,\alpha} = \max\{l : P_{\theta_0}(X \leq l) \leq \alpha/2\}$ bzw. $c_{2,\alpha} = \min\{l : P_{\theta_0}(X \geq l) \leq \alpha/2\}$. Die beiden Testentscheidungen

$$\text{Ablehnung von } H_0, \text{ falls } p_{\mathrm{B}'}(k) \leq \alpha$$

und

$$\text{Ablehnung von } H_0, \text{ falls } k \in K_{\mathrm{B}'}(\alpha)$$

sind nur im Fall des p-Wertes in der Version 1 identisch. Für die Versionen 2 oder 3 gilt dies im Allgemeinen nicht mehr!

Zahlenbeispiel: (i) Sei $\theta_0 = 0.25$ und $n = 20$.

Im Fall $k = 1$ ergibt Version 1

$$p_{\mathrm{B}'}(1) = 2 \cdot P_{0.25}(\tilde{X} \leq 1) = 0.049$$

Version 2

$$p_{\mathrm{B}'}(1) = P_{0.25}(\tilde{X} \leq 1) + P_{0.25}(\tilde{X} \geq 9) = 0.065$$

und Version 3

$$p_{\mathrm{B}'}(1) = P_{0.25}(\tilde{X} \leq 1) + P_{0.25}(\tilde{X} \geq 10) = 0.038$$

Im Fall $k = 9$ ergibt Version 1

$$p_{\mathrm{B}'}(9) = 2 \cdot P_{0.25}(\tilde{X} \geq 9) = 0.082$$

Version 2 und 3 stimmen überein:

$$p_{\mathrm{B}'}(9) = P_{0.25}(\tilde{X} \geq 9) + P_{0.25}(\tilde{X} \leq 1) = 0.065$$

Im Fall $k = 10$ ergibt Version 1

$$p_{\mathrm{B}'}(10) = 2 \cdot P_{0.25}(\tilde{X} \geq 10) = 0.028$$

Die Versionen 2 und 3 stimmen wieder überein:

$$p_{\mathrm{B}'}(10) = P_{0.25}(\tilde{X} \geq 10) + P_{0.25}(\tilde{X} = 0) = 0.017$$

Wie sieht der kritische Bereich zum Signifikanzniveau $\alpha = 0.05$ aus? Der untere kritische Wert ist $c_{1,\alpha} = 1$, der obere kritische Wert ist $c_{2,\alpha} = 10$. Damit ist der kritische Bereich $\{0,1\} \cup \{10,\ldots,20\}$. Wir sehen also, dass bei $k = 1$ der p-Wert nach Version 2 von 0.065 zu keiner Ablehnung von H_0 führt. Nach der Neyman-Pearson-Methode würden wir ablehnen, da 1 im kritischen Bereich liegt.

(ii) Sei nun $\theta_0 = 0.2$ und $n = 20$. Im Fall $k = 8$ ergibt Version 1

$$p_{\mathrm{B}'}(8) = 2 \cdot P_{0.2}(\tilde{X} \geq 8) = 0.064$$

Version 2 und 3 stimmen überein:

$$p_{\mathrm{B}'}(8) = P_{0.2}(\tilde{X} \geq 8) + P_{0.25}(\tilde{X} = 0) = 0.044$$

Zum Signifikanzniveau $\alpha = 0.05$ ist $\{0\} \cup \{9,\ldots,20\}$ der kritische Bereich. Auch hier sehen wir: Verwendet man die Version 2 oder 3, so führt der p-Wert von 0.044 zu einer Ablehnung von H_0, während es nach der Neyman-Pearson-Methode zu keiner Ablehnung von H_0 kommt, da 8 nicht im kritischen Bereich liegt.

Version 1 ist wohl die gebräuchlichste (Sheskin 2011, Seite 313) und entspricht der Festlegung im stetigen Fall (siehe Bemerkung 1).

3.6 Approximativer Binomialtest

Für große Stichprobenumfänge verwendet man die Prüfgröße

$$Z = \frac{X - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}},$$

die nach dem zentralen Grenzwertsatz asymptotisch $N(0,1)$-verteilt ist. Ist k eine Realisierung von X und schreiben wir

$$z_k = \frac{k - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

so gilt $p_{\mathrm{L}'}(k) \approx \Phi(z_k)$, $p_{\mathrm{R}'}(k) \approx 1 - \Phi(z_k)$, $p_{\mathrm{B}'}(k) \approx 2 \cdot (1 - \Phi(|z_k|))$.

4 Aspekte der Diskussion um p-Werte

In diesem Abschnitt werden Probleme bei der Verwendung von Software, p-Wert versus Signifikanzniveau α und Fragen des Schulunterrichts behandelt.

4.1 Der p-Wert bei statistischer Software

Gängige statistische Softwarepakete wie R, SAS und SPSS weisen den p-Wert automatisch aus. Wir wollen auf zwei Punkte aufmerksam machen, die es dabei zu beachten gilt:

1. *Einseitiger p-Wert*: Darunter verstehen die Programme (z. B. SAS) das Minimum vom rechtsseitigen und linksseitigen p-Wert. In diesem Fall muss diese Zahl nicht der p-Wert von dem tatsächlich interessierenden Testproblem sein. Liegt also im Binomialmodell das rechtsseitige Testproblem (R') zugrunde und gilt $p_{\mathrm{L}'}(k) = \min\{p_{\mathrm{L}'}(k), p_{\mathrm{R}'}(k)\}$, so ist die im Programm-Output angegebene Zahl $p_{\mathrm{L}'}(k)$ der linksseitige p-Wert und nicht der rechtsseitige p-Wert, welchen man dann so bestimmt: $p_{\mathrm{R}'}(k) = 1 - p_{\mathrm{L}'}(k) + P_{\theta_0}(\tilde{X} = k)$.

Man hat also bei der Interpretation des Outputs zusätzlich zu berücksichtigen, ob der Prüfgrößenwert k überhaupt extrem im Sinne der Alternative (R') ist. Eine Entscheidung für die Alternative $H_1 : \theta > \theta_0$ ist nur dann sinnvoll, wenn k rechts vom Erwartungswert $n\theta_0$ liegt. Bei anderen Programmen wie R gibt es die Option, links- oder rechtsseitig zu testen.

2. *Zweiseitiger p-Wert*: Im nichtsymmetrischen Fall des Binomialtests verwenden statistische Softwarepakete unterschiedliche Versionen. Beispielsweise verwenden SAS und SPSS die Version 1, R verwendet Version 3. Wird nur ein zweiseitiger p-Wert in der Version 1 angegeben (z. B. SPSS), so ist dieser durch 2 zu teilen, um den einseitigen p-Wert zu erhalten. Bei der Interpretation dieser Zahl ist Punkt 1 zu beachten.

Bei Verwendung von Software, die den p-Wert beim Binomialtest nicht berechnen (wie z. B. Excel), bleibt nur, den p-Wert über die (kumulierten) Einzelwahrscheinlichkeiten zu bestimmen.

4.2 Diskussion: p vs α?

Die bisherigen Ausführungen suggerieren, dass es keine wesentlichen Unterschiede zwischen der Vorgabe eines Signifikanzniveaus α (Neyman-Pearson-Methode) und der p-Wert-Methode (Fisher) gibt. Eine wichtige Gemeinsamkeit ist, dass beide Verfahren von der Gültigkeitsannahme der Nullhypothese ausgehen (kein Bayes'scher Ansatz!). Aber es gibt Unterschiede zwischen diesen beiden Ansätzen. Wir möchten hier nur auf einen Punkt aufmerksam machen.

Nach der Neyman-Pearson-Methode stehen sich zwei konkurrierende Hypothesen – Nullhypothese H_0 und Alternativhypothese H_1 – gegenüber mit den zentralen Begriffen Fehler 1. Art (α-Fehler) und Fehler 2. Art (β-Fehler). Fisher hat (die explizite Formulierung von) Alternativen abgelehnt und ein Verwerfen von H_0 bedeutet (noch) nicht, dass man sich für eine alternative Hypothese entscheiden soll. Eigentlich kaum zu glauben, wenn man an den p-Wert im Sinne von Fisher denkt: Ein Signifikanztest ist ein Verfahren, das eine Wahrscheinlichkeit (unter H_0) berechnet, das beobachtete Ergebnis oder noch extremere Ergebnisse zu erzielen. Was extrem bedeutet, wird letztlich durch die Alternativhypothese bestimmt, denn sie legt die *Richtung* der Abweichungen von H_0 fest (Fisher hat daher zumindest implizit an Alternativen gedacht).

Prüft man in einem Normalverteilungsmodell den Mittelwert, dann verwendet man bei bekannter Varianz den Gauß-Test (bei unbekannter Varianz ist es der t-Test), sind dagegen Aussagen über die Varianz interessant, dann verwendet man den χ^2-Test auf Varianz.

Der p-Wert wird sich auf denjenigen Test beziehen, der die Alternative möglichst gut entdeckt, der also eine möglichst hohe *Macht* (Power), *Schärfe* hat. So ist z. B. der Gauß-Test unter den getroffenen Verteilungsannahmen ein bester unverfälschter Niveau-α-Test (siehe z. B. Georgii 2009, Kapitel 10). Bezüglich der wichtigsten in der Praxis verwendeten Testverfahren sei auf das Standardwerk von Sheskin 2011 verwiesen.

Über den Konflikt „p-Wert gegen festes Niveau", der die Auseinandersetzungen zwischen den Begründern der heutigen Testtheorie R. A. Fisher (1890-1962) auf der einen Seite und J. Neyman (1894-1981) und E.S. Pearson (1895-1980) auf der anderen Seite widerspiegelt, sei auf den Artikel von Hubbard und Bayarri 2003 mit den sich anschließenden Diskussionsbeiträgen von Berk 2003 und Carlton 2003 verwiesen. Lehmann 1993 beleuchtet das Thema (erfreulicherweise) mehr aus „statistischer" als aus „philosophischer" Sicht. Eine lesenswerte historische Einordnung gibt Stute 1989, siehe auch Sheskin 2011, S. 68-74.

4.3 Der p-Wert im Schulunterricht

Zunächst darf das Konzept des p-Wertes nicht dazu (ver)führen, auf das rational-mathematische Konzept der Neyman-Pearson-Testtheorie zu verzichten. Dies bedeutet: Das Testen von Hypothesen darf nicht beim Signifikanztest (Niveau-α-Test) stehen bleiben, wo es nur um den α-Fehler geht. Dies wäre nur eine Seite der Medaille. Man muss den β-Fehler bzw. die Macht, die Power ($= 1 - \beta$) eines Tests mit ins Spiel bringen. Ohne die Macht anzusprechen ist es schwierig, folgende Fragen zu thematisieren:

- Die Wahl von $\alpha = 0.05$ als Kompromisslösung ist erst durch die Gegenläufigkeit der beiden Fehlerwahrscheinlichkeiten ($\alpha \downarrow \Rightarrow \beta \uparrow$ bzw. $\alpha \uparrow \Rightarrow \beta \downarrow$) zu begründen (warum nicht $\alpha = 10^{-6}$?)
- Die Bedeutung des Stichprobenumfangs bleibt unklar. Welchen Einfluss hat der Stichprobenumfang auf die Power bzw. den β-Fehler?
- Welche Rolle spielen Unterschiede von praktischer Relevanz (Effektgrößen)?
- Wie groß muss der Stichprobenumfang mindestens sein, um einen Effekt mit einer gewissen Wahrscheinlicheit zu entdecken?

Für den Zweistichproben-Gauß-Test siehe dazu auch Börgens 2014. Wenn die „Philosophie" der Neyman-Pearson-Methode verstanden worden ist, ihre (zugegebenermaßen nicht einfach zu verstehenden) Begriffsbildungen und Sprechweisen geklärt und Fehlinterpretationen ausgeschlossen sind (an dieser Stelle sei noch einmal auf Henze 2013, Kap. 29 verwiesen), kann der p-Wert eingeführt werden (zuerst links- bzw. rechtsseitig, dann beidseitig). Dieser soll als eine (schnelle) Entscheidungshilfe für den Niveau-α-Test gesehen werden (man erspart sich den Vergleich mit Quantilen aus Tabellenwerken): Eine Nullhypothese ist zu verwerfen, falls der p-Wert $\leq \alpha$ ist.

Im Diskreten wird bei zweiseitiger Fragestellung empfohlen, den p-Wert in der Version 1 einzuführen,

damit auch in diesem Fall die Äquivalenz der beiden Entscheidungsvorschriften (zum Niveau α) „Ablehnung von H_0, falls p-Wert $\leq \alpha$" und „Ablehnung von H_0, falls der Prüfgrößenwert im kritischen Bereich K_α liegt" gewährleistet ist. Das „Ausweichen" auf den approximativen Binomialtest ist weniger zu empfehlen.

Egal, ob der p-Wert als Wahrscheinlichkeit oder als Grenzniveau interpretiert wird: Wichtig ist es zu betonen, was der p-Wert *nicht* ist: Wahrscheinlichkeit für die Richtigkeit einer Nullhypothese bzw. Wahrscheinlichkeit für den Fehler 1. Art.

Beim p-Wert hat die Bayes-Statistik nichts zu suchen. Ein Grund mehr bei der Einführung des Hypothesentests nach der Neyman-Pearson-Methode die Bayes-Statistik außen vor zu lassen. In diesem Zusammenhang sei an Diepgen 2002 erinnert. Erst danach (wenn überhaupt) sollte das Konzept der Bayes-Statistik vorgestellt werden. Dabei ist zu beachten, dass die *frequentistische* Sichtweise und die *Bayes'sche* Sichtweise zwei völlig verschiedene Konzepte sind, deren Wahrscheinlichkeitsaussagen sich überhaupt nicht sinnvoll miteinander vergleichen lassen (man vergleicht Äpfel mit Birnen). Wie schreibt Carlton (2003) so treffend:

> „A 5% cutoff means that 5% of true hypotheses are rejected, not that 5% of rejected hypotheses are true."

Und wer die Bayes-Statistik unbedingt im Unterricht behandeln will, sei noch einmal daran erinnert: Der p-Wert ist keine inverse Wahrscheinlichkeit (a-posteriori-Wahrscheinlichkeit).

5 Schlussbemerkungen

In der Literatur findet man gelegentlich die Auffassung, dass der p-Wert ein Maß für die Verträglichkeit (measure of evidence) von Daten und Nullhypothese ist (siehe z. B. Stahel 2008, Kap. 8). Es gibt aber auch Kritik an einer solchen Auffassung (Schervish, 1996).

Auf eine Gefahr wollen wird abschließend hinweisen und zwar auf die Gefahr, dass das Signifikanzniveau an den p-Wert angepasst wird. Wenn der p-Wert angegeben wird, so hat im Prinzip jeder das Recht zu dem Niveau zu testen, das er für geeignet hält. Aber diese Meinungsfreiheit untergräbt den wahren Sinn des statistischen Testens. Angenommen, der p-Wert ist 0.042. Möchte man ein signifikantes Testresultat, dann wählt man (im Nachhinein!) $\alpha = 0.05$. Soll die Nullhypothese nicht abgelehnt werden (z. B. aus bestimmten Interessensgründen), so wählt man $\alpha = 0.01$. Daher ist die Angabe einer oberen Schranke α für den p-Wert – und zwar bevor man den p-Wert vom Computer ausrechnen lässt – unentbehrlich.

Der p-Wert hat seine Tücken. Daher ist es wichtig, ein Verständnis für diesen Begriff zu entwickeln. Auch deswegen, damit die Statistik nicht zu einer schwarzen Kiste wird. Computerprogramme berechnen nämlich den p-Wert aus den Daten per Knopfdruck.

Dem Anwender soll es nicht so gehen wie wohl den meisten Lesern des Science Fiction Romans *Per Anhalter durch die Galaxie* von Douglas Adams. Hier hatte der Supercomputer *Deep Thought* die Frage nach dem Leben, dem Universum und dem ganzen Rest nach über sieben Millionen Jahren Rechenzeit mit der Zahl „42" beantwortet. Unbefriedigend (die Antwort, weniger die Rechenzeit). Schade, dass *Deep Thought* nicht die Zahl 0.042 ausgewiesen hat. Dann hätte sich aus statistischer Sicht eine Interpretationsmöglichkeit angeboten. Und hier wird deutlich: Ob das Ergebnis 0.042 als signifikant einzustufen ist, dazu hätte man sich auf ein Niveau α einigen müssen, bevor *Deep Thought* diesen Wert ausgespuckt hätte. Zeit genug zur Einigung wäre jedenfalls gewesen.

Literatur

Berk, K. N. (2003): Discussion of a paper by Hubbard and Bayarri (Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing). In : *The American Statistician* 57(3), S. 178-179.

Börgens, M. (2014): Die Bedeutung des β-Risikos. In: *Stochastik in der Schule* 34(1), S. 8-12.

Carlton, M. A. (2003): Discussion of a paper by Hubbard and Bayarri (Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing). In : *The American Statistician* 57(3), S. 179-181.

Diepgen, R. (2002): $P(H|D)$ versus $P(D|H_0)$? Wie man das Testen von Hypothesen - lieber doch nicht - einführen sollte. In: *Stochastik in der Schule* 22(3), S. 34-38.

Falk, M., Becker, R. und Marohn, F. (2004): Angewandte Statistik. Berlin-Heidelberg: Springer.

Falk, M., Hain, J., Marohn, F., Fischer, H. und Michel, R. (2014): Statistik in Theorie und Praxis. Mathematik für das Lehramt, Springer-Spektrum, Berlin-Heidelberg.

Freund, J.E.; Perles, B.M. (1996): Einige Beobachtungen zur Definition von p-Werten: *Stochastik in der Schule* 16(2), S. 36-38.
Georgii, H.-O. (2009): Stochastik. 4. Auflage. Berlin: de Gruyter.
Henze, N. (2013): Stochastik für Einsteiger. 10. Auflage. Wiesbaden: Springer Spektrum.
Hubbard, R. und Bayarri, M. J. (2003): Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. In: *The American Statistician 57, S. 171-178*.
Krengel, U. (2005): Einführung in die Wahrscheinlichkeitstheorie und Statistik. 8. Auflage. Wiesbaden: Vieweg.
Lehmann, E. L (1993): The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two?: In: *Journal of the American Statistical Association* 88 (424), S. 1242-1249.
Rudolf, M. und Kuhlisch, W. (2008): Biostatistik. München: Pearson Studium.
Schervish, M. J. (1996): P values: What they are and what they are not. In: *The American Statistician* 50(3), S. 203-206.
Sheskin, D. J. (2011): Handbook of Parametric and Nonparametric Statistical Procedures. Fifth Edition. Boca Raton: Chapman & Hall.
Stahel, W. A. (2008): Statistische Datenanalyse. 5. Auflage. Wiesbaden: Vieweg.
Stute, W. (1989): Der historische Streit zwischen R. A. Fisher und J. Neyman oder: Ein Sittengemälde aus der Blütezeit der englischen Schule für Statistik. In: *Mathematische Semesterberichte* 36(1), S. 61-84.
Wengenroth, J. (2008): Wahrscheinlichkeitstheorie. Berlin: de Gruyter.

Anschrift des Verfassers
Frank Marohn
Institut für Mathematik
Universität Würzburg
Emil-Fischer-Str. 30
97074 Würzburg
marohn@mathematik.uni-wuerzburg.de

Rund um den Variationskoeffizienten in einführenden Statistikkursen[1]

DAVID TRAFIMOW, USA

[1] Das Original erschien in Teaching Statistics (Volume 36, Number 3, Autumn 2014; S. 81–82). Originaltitel: On teaching about the coefficient of variation in introductory statistics courses
Übersetzung: R. VEHLING

Zusammenfassung: *Die Standardabweichung wird mithilfe des Variationskoeffizienten auf den Mittelwert bezogen. Lehrende in Statistikkursen können durch dieses Vorgehen den Lernenden die Standardabweichung verständlicher vermitteln.*

1 Einleitung

In einführenden Statistikkursen gehören das (arithmetische) Mittel sowie die (empirische) Standardabweichung zum Pflichtprogramm. Lernende kommen mit einigen Kenntnissen über den Mittelwert in Statistikkurse. Der Mittelwert ist einfach zu berechnen und Lernende sind damit vertraut, da diese Kenngröße bei der Berechnung des Notendurchschnitts sehr oft eingesetzt wird. Im Gegensatz dazu stellt die Standardabweichung eine größere Herausforderung in einführenden Statistikkursen dar. Die Berechnung der Standardabweichung ist etwas schwieriger als beim Mittelwert. Aber das ist nicht die eigentliche Herausforderung. Die wahre Herausforderung besteht darin, dass die Lernenden nicht mit der Standardabweichung vertraut sind und wenige Beispiele haben, die sie auf diese Kenngröße beziehen können.

2 Mögliches Vorgehen

Zurück zum Mittelwert. Falls einem Lernenden mitgeteilt wird, dass die durchschnittliche Leistung bei einer Prüfung den Wert 0,85 ergeben hat, kann er diesen Wert auf seine eigene Leistung beziehen. In vielen Klassen, in denen die Standardskala benutzt wird (0,9; 0,8; 0,7; und 0,6 als jeweilige Schwellenwerte für A, B, C und D), folgt daraus, dass die Durchschnittswerte sich im mittleren Bereich von „B“ befinden. Lernende können das fast automatisch sehen. Falls einem Lernenden mitgeteilt wird, dass die durchschnittliche Leistung bei einer Prüfung den Wert 0,55 hat, wird er wahrscheinlich daraus folgern, dass die Prüfung schwer war, zumindest verglichen mit den üblichen Standards.

Im Gegensatz dazu nehmen wir nun an, einem Lernenden wird mitgeteilt, dass die Standardabweichung bei einer Prüfung 0,05 beträgt. Ist dies ein großer oder ein kleiner Wert? Was bedeutet 0,05? Oder nehmen wir den Wert 0,20. Ist dieser Wert groß oder klein? Was bedeutet 0,20?

Da Lernende eine viel bessere intuitive Vorstellung vom Mittelwert als von der Standardabweichung haben, macht es Sinn – falls es möglich ist – einen Weg zu finden, diese beiden Begriffe für die Lernenden zu vernetzen. Natürlich ist dies möglich, und zwar mithilfe des so genannten Variationskoeffizienten C_V, der das Verhältnis aus der Standardabweichung s und dem Mittelwert $\bar{x}$ beschreibt $\left(C_V = \frac{s}{\bar{x}}\right)$.

Betrachten wir zum Beispiel unter Berücksichtigung der Beschreibenden Statistik einen Test, der von Dr. Nasty und einen anderen Test, der von Dr. Nice konstruiert wurde. Für den Test von Dr. Nasty sind der Mittelwert 0,30 und die Standardabweichung 0,10. Für den Test von Dr. Nice sind der Mittelwert 0,90 und die Standardabweichung 0,15. Ohne Berücksichtigung der Mittelwerte ist es schwierig zu erkennen, ob 0,10 und 0,15 große oder kleine Werte für die entsprechenden Standardabweichungen sind, obwohl es auf der Hand liegt, dass 0,15 eine größere Zahl als 0,10 ist. So könnte man versucht sein zu schließen, dass der Test von Dr. Nasty weniger gut zwischen den Lernenden unterscheidet als der Test von Dr. Nice. Wird andererseits der Variationskoeffizient berücksichtigt, so ergeben sich für die beiden Tests von Dr. Nasty und Dr. Nice die Werte 0,33 beziehungsweise 0,17. Bezogen auf die Größe des Mittelwertes ist der Test von Dr. Nasty differenzierter als der Test von Dr. Nice. Für die Lernenden ist es leicht einsehbar, ob eine Standardabweichung groß oder klein ist. Der Vergleich zweier Standardabweichungen kann aber von den entsprechenden Mittelwerten abhängig sein.

Betrachten wir ein weiteres Beispiel: Die Standardabweichungen der Massen der acht Planeten (Pluto wird nicht mehr als Planet angesehen) haben abhängig von der benutzten Maßeinheit Kilogramm oder Pfund die Maßzahlen 2,74E + 12 beziehungsweise 6,04E + 12. Obwohl die beiden Standardabweichungen zahlenmäßig sehr unterschiedlich sind, ergibt sich für den jeweiligen Variationskoeffizienten der gleiche Wert (1,85), wenn zusätzlich die unterschiedlichen Maßzahlen der Massen (1,48E + 12 bzw. 3,26E + 12) berücksichtigt werden. Auch dieses weitere Beispiel macht deutlich, wie der Variationskoeffizient dazu dienen kann, den Begriff der Standardabweichung besser zu verstehen.

Der Variationskoeffizient kann auch bei den Geschwindigkeiten der Planeten angewendet werden. Die Standardabweichung beträgt 21,404 km/s beziehungsweise 21404 m/s. Somit ist die letztgenannte Standardabweichung 1000 Mal größer als die Erstgenannte. Obwohl sich die Mittelwerte im gleichen Maße unterscheiden, sind die Variationskoeffizienten gleich, nämlich 0,669.

3 Fazit

Zusammenfassend kann das Folgende gesagt werden: Der Mittelwert und die Standardabweichung sind wichtige Kenngrößen. Für die Lernenden ist es schwierig, den Begriff der Standardabweichung zu verstehen. Lehrende sollten in Statistikkursen, die der Beschreibenden Statistik gewidmet sind, das Konzept des Variationskoeffizienten in ihren Kursen mit einbeziehen.

Auf diese Weise haben Lernende eine größere Chance, die Bedeutung der Standardabweichung zu verstehen, da der Variationskoeffizient einen expliziten Zusammenhang zwischen der Standardabweichung und dem Mittelwert herstellt.

Literatur

Fisher, R. A. (1925). Statistical Methods for Research Workers on the Development of the Science of Statistics. Edingburgh: Oliver and Boyd.

Yadav, R., Upadhyaya, L. N., Singh, H. P., Chatterjee, S. (2013). A general procedure of estimating the population variance when coefficient of variation of an auxiliary variable is known in sample surveys. *Quality & Quantity: International Journal of Methodology*, **47**(4), 2331–2339. DOI: 10.1007/s11135-012-9659-6

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, **46**(253), 19–34. DOI: 10.2307/2280090

Anschrift der Verfasser

David Trafimow
Departement of Psychology, New Mexico State University, Las Cruces, NM, USA
dtrafimo@nmsu.edu

Rezension

Haller, Rudolf & Barth, Friedrich: Berühmte Aufgaben der Stochastik

De Gruyter – Oldenbourg; München 2014

JÖRG MEYER, HAMELN

Generationen von Schülerinnen und Schülern haben Stochastik nach dem gleichnamigen Schulbuch von F. Barth und R. Haller gelernt; die Autoren haben sich damit um die Förderung der Stochastik sehr verdient gemacht. Erfreulicherweise haben sie das nicht zum Anlass genommen, sich auf ihren Lorbeeren auszuruhen (wie auch schon aus der Fülle der Aufsätze beider Autoren in dieser Zeitschrift deutlich wird), sondern haben mit großem Fleiß und fast noch größerer Akribie in vielerlei Quellen geforscht, um die Genese und die ersten Lösungsansätze (samt moderner Behandlung) von vielerlei Aufgaben aus der Stochastik darzustellen. Das Resultat ist ein auch handwerklich schön gemachtes Buch mit nur wenigen Druckfehlern.

Das Adjektiv im Titel „Berühmte Aufgaben" stellt eine Untertreibung dar: Nicht alle im Buch enthaltenen Aufgaben sind nach meiner Kenntnis „berühmt"; viele jedoch sind zumindestens interessant.

Das Substantiv „Aufgaben" hingegen ist treffend gewählt: Die Autoren verzichten weitgehend darauf, die Entwicklung der Aufgabenkontexte im Laufe der Zeit darzustellen (d. h. wie sich eine Aufgabe gewandelt oder erweitert hat, zu welchen Begriffsbildungen oder Theorien sie Anlass gegeben hat), sondern lassen den Leser selber entdecken, wie viel „Musik" in vielen alten Aufgaben steckt. Diesem Konzept entspricht die sich an den Jahreszahlen der Erstpublikationen orientierende Anordnung; will man hingegen wissen, wann etwa der Erwartungswert zum ersten Mal verwendet wurde, muss man im Buch etwas suchen. Für eine weitere Auflage sollte das Sachregister ausführlicher ausfallen.

Naturgemäß gehören die Aufgaben bis etwa zum 17. Jahrhundert vor allem zur Kombinatorik.

Das tatsächlich sehr berühmte problème des parties (ursprünglich von 1494) wird ausführlich mit unterschiedlichen Lösungsvarianten geschildert. Diverse Probleme von Huygens (und später erwartungsgemäß bei Jakob Bernoulli) befassen sich mit der Binomial- bzw. mit der hypergeometrischen Verteilung, aber auch mit Sterbetafeln und dem Median (1669). Der Unterschied zwischen arithmetischem Mittel und Median kommt auch bei Cardano und Jakob Bernoulli zum Tragen; letzterer hat auch wohl als erster mit dem Erwartungswert operiert (der im Buch zwar schon vorher vorkam, aber, wenn ich es richtig gesehen habe, immer nur in den modernen Lösungen).

Offenbar war Jakob Bernoulli auch der erste, in dessen Lösung eines Problems von Montmort die Verwendung von erzeugenden Funktionen (natürlich, ohne sie explizit zu verwenden oder sie gar so zu benennen) hineininterpretiert (oder herausgelesen) werden kann.

Das berühmte Petersburg-Paradoxon von Nikolaus Bernoulli mit seinen unterschiedlichen Varianten wird mit diversen historischen Lösungsansätzen ausführlich präsentiert. Schon in dieser Zeitschrift (Barth/Haller (2011)) haben die Autoren die Behandlung des Problems, mit welcher Wahrscheinlichkeit morgen die Sonne aufgeht, in historischer Perspektive behandelt – ein Problem, dessen philosophische Tragweite (über Hume zu Kant) bedeutend größer ist als die mathematische. Die damit verbundene „Folgeregel" und die Formel von Bayes werden ausführlich in ihren historischen Ansätzen geschildert. Das ebenfalls in dieser Zeitschrift schon dargestellte Casanova-Problem (Barth/Haller (2012a)), wonach bei der Augensumme zweier Würfel gerade Zahlen häufiger vorkommen sollen als ungerade, lässt sich schon mit elementaren Mitteln einsichtig behandeln und wird von den Autoren angemessen verallgemeinert.

Die im Vorwort genannte Beschränkung, „zu zeigen, wie die Mathematiker mit den Hilfsmitteln ihrer Zeit" ihre Probleme lösten, wird mit Augenmaß eingehalten und nicht blind befolgt. Daher kann man der Ansicht sein, dass die Paradoxien bei den Wahlverfahren nach Borda und Condorcet sowie der damit zusammenhängende Unmöglichkeitssatz von Arrow eine wesentlich ausführlichere Behandlung (vgl. etwa Meyer (1995 und 1998)) verdient hätten.

Ab der zweiten Hälfte des 19. Jahrhunderts werden die Probleme deutlich vielfältiger. Man erfährt, dass sich auch Richard Dedekind mit Stochastik befasst hat und dabei auch gleich ein Urnenparadoxon gefunden hat, das zumindestens mir vorher nicht be-

kannt war. Hermann Laurent hat das Sinusprodukt erfolgreich auf stochastische Fragestellungen angewendet (die Laurent-Reihen sind nach P. A. Laurent benannt). Natürlich gab es viele Probleme rund um die Bayes-Formel (etwa Bertrands Drei-Kästchen-Problem). Größeren Raum nimmt das berühmte Geburtstagsproblem ein, das offenbar bei von Mises zum ersten Mal auftritt, zusammen mit naheliegenden Verallgemeinerungen (vgl. auch Barth/Haller (2012b und 2013)).

Nicht jeder Leser wird alle Probleme gleich interessant finden, aber wohl jeder Leser wird mehrere Probleme finden, die ihn ansprechen. Bei mir war es vor allem das Konzept des Begünstigens nach Chung (1942), das mich zu eigenen Überlegungen angeregt hat (danach erinnerte ich mich an die schöne Analyse von Borovcnik (1992), Kap. 3.2). „Begünstigt" und „benachteiligt" sind symmetrische Relationen; sie sollten mit symmetrischen Symbolen (wie „↑" und „↓") gekennzeichnet werden. In diesem Unterabschnitt findet sich, was man bei diesen Autoren gar nicht erwartet hätte, ein falscher Beweis (zur Nichttransitivität von „benachteiligt"; das Zahlentripel $\{-3; 2; 10\}$ ist ein einfaches Gegenbeispiel für die Nichttransitivität). Dass die Autoren Chungs aufwändiges Gegenbeispiel für die Nichttransitivität von „begünstigt" zitieren, zeigt die historische Sorgfalt ($\{-10; -3; 2\}$ hätte es auch getan).

Dass Rekorde mit Potenzsummen und insofern mit Bernoulli-Zahlen zusammenhängen, wird nicht jeder Leser vermutet haben. Haller und Barth beschreiben die Genese bei Joseph Bertrand (der u. a. durch das nach ihm benannte Drei-Kästchen-„Paradoxon" berühmt wurde) sowie die Behandlung der Rekorde bei Arthur Engel. Dieser Ansatz wird bei Henze (2008) ausgebaut.

Dass das berühmte und vielfältig behandelte „Problem des anderen Kindes" in anderer (zunächst viel komplizierter) Formulierung von Erwin Schrödinger stammt: Auch so etwas erfährt man bei Haller und Barth.

Das Werk schließt mit dem Simpson-Paradoxon und dem pairwise-worst-best paradox von Blyth, in dem allerdings deutlich mehr steckt, als von den Autoren angedeutet wird (vgl. etwa Meyer (1995)).

Manche Probleme wiederholen sich – nicht aber die Lösungswege! So wird die Summe der Quadrate von Binomialkoeffizienten von mehreren Autoren recht unterschiedlich angegangen.

Die profunde humanistische Bildung der Autoren des vorliegenden Buches zeigt sich auch in Kleinigkeiten: Der Leser erfährt beispielsweise, warum Voltaire sich so genannt hat (S. 70), inwiefern Caesars „alea iacta est" eine irreführende Übersetzung ist (S. 96: Die Würfel sind nicht schon gefallen, sondern sind erst hochgeworfen), woher das „Korollar" stammt (S. 142), wann das Lotto in seiner heutigen Form eingeführt wurde (S. 145) und vieles mehr; man bekommt auch einen Eindruck von der Intensität, mit der sich etwa Pascal stochastischen Problemen widmete.

Insgesamt handelt es sich um ein sehr schönes Buch, das zu vielen eigenen Überlegungen anregt!

Literatur

Barth, F./Haller, R. (2011): Geht die Sonne morgen wieder auf?. In: *Stochastik in der Schule* 31 (2); S. 6–17.

Barth, F./Haller, R. (2012a): Numero deus impare gaudet. In: *Stochastik in der Schule* 32 (1); S. 15–20.

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

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

Borovcnik, M. (1992): Stochastik im Wechselspiel von Intuitionen und Mathematik. Mannheim: BI Wissenschaftsverlag.

Henze, N. (2008): Rekorde. In: *Der Mathematikunterricht* 54 (1), S. 16–23.

Meyer, J. (1995): Einfache Paradoxien der beschreibenden Statistik. In: *Stochastik in der Schule* 15 (2), S. 27–50.

Meyer, J. (1998): Paradoxien bei direkten Wahlen. In: *Mathematik lehren* 88, S. 50–54.

Einladung zur Herbsttagung 2015 des Arbeitskreises Stochastik

KATJA KRÜGER, PADERBORN

Liebe Kolleginnen und Kollegen,

hiermit möchten wir Sie ganz herzlich zur diesjährigen Herbsttagung des Arbeitskreises Stochastik vom Freitag, 20.11.2015 (abends) bis zum Sonntag, 22.11.2015 (mittags), im InVia-Gästehaus in Paderborn einladen. Das Thema der Tagung lautet dieses Jahr **„Digitale Medien im Stochastikunterricht“**. Dabei sollen nicht nur etablierte Werkzeuge wie GTR, CAS oder geeignete PC-Software, sondern auch die neuen Möglichkeiten und Grenzen von Apps oder Videos für das Lernen von Stochastik in den Blick genommen werden.

Wie bei den vergangenen Herbsttagungen werden auch diesmal Vorträge eingeworben, durch die das Tagungsthema aus verschiedenen Perspektiven beleuchtet werden soll. Die themenspezifischen Vorträge sind gegenüber anderen (freien) Vorträgen hinsichtlich der Vortragsdauer und der Diskussionszeit verlängert. Über die themenspezifischen Vorträge hinaus gibt es traditionell die Möglichkeit, freie Vorträge (30 min Vortrag & 15 min Diskussion) anzubieten. Solche Angebote senden Sie bitte bis zum 31.08.2015 an die Mail-Adresse kakruege@math.upb.de.

Die Kosten für die Herbsttagung (Tagungsgebühr + Vollverpflegung + Kaffee + Übernachtung) werden ca. 155 Euro betragen. Die Anmeldung erfolgt per E-Mail an die oben angegebene Adresse. Bitte melden Sie sich bis zum 30.09.2015 an. Der konkrete Programmablauf, die genauen Kosten sowie alle weiteren Informationen werden wie üblich auf der Homepage des Arbeitskreises veröffentlicht und per Rundmail bekannt gegeben.

http://www.mathematik.uni-dortmund.de/ak-stoch/

Für die Aufnahme in den Mail-Verteiler des Arbeitskreises schicken Sie bitte bei Bedarf einen entsprechenden Hinweis an die oben genannte E-Mail-Adresse.

Wir freuen uns auf eine anregende Tagung in Paderborn.

Katja Krüger und Philipp Ullmann

(Sprecher des Arbeitskreises Stochastik)

Einladung zur Mitgliederversammlung des Vereins zur Förderung des schulischen Stochastikunterrichts e. V.

Hiermit wird eingeladen zur Mitgliederversammlung des Vereins zur Förderung des schulischen Stochastikunterrichts e. V. anlässlich der Herbsttagung 2015 des Arbeitskreises „Stochastik in der Schule“ in der GDM.

Den Raum und die Zeit entnehmen Sie bitte dem schwarzen Brett im Tagungsbüro.

Da ich zum Jahresende aus dem Vorstand ausscheiden möchte, muss nachgewählt werden.

Tagesordnung:

1. Eröffnung und Beschlussfassung über die Tagesordnung
2. Vorstandsnachwahl
3. Verschiedenes

gez. Jörg Meyer (Schriftführer)

Bibliographische Rundschau

GERHARD KÖNIG, KARLSRUHE

Vorbemerkung: *Die hier nachgewiesenen Veröffentlichungen sind alphabetisch nach dem Erstautor angeordnet. Ein Kurzreferat versucht, die wesentlichen Inhalte der nachgewiesenen Zeitschriftenaufsätze und Bücher wiederzugeben.*

Daniel Bättig: Angewandte Datenanalyse: Der Bayes'sche Weg. Berlin; Heidelberg: Springer Spektrum, 2015

Angewandte Datenanalyse, Bayes'sche Statistik und moderne Simulationsmethoden mit dem Computer helfen, nicht direkt messbare Größen zu bestimmen und Prognosen zu zukünftigen Werten von unsicheren Größen zu berechnen. Wie dabei vorgegangen werden kann, von der systematischen Sammlung von Daten, von der Frage wie Unsicherheit mit Wahrscheinlichkeiten quantifiziert werden kann, bis hin zu Regressionsmodellen, spannt das Buch den Bogen. Durch seinen systematischen Aufbau mit zahlreichen Beispielen aus der Praxis und seine in vielen Kursen erprobte Didaktik ist das Buch ideal für Studierende in den angewandten Wissenschaften wie Ingenieur-, Natur- und Wirtschaftswissenschaften geeignet. (Autorreferat)

Michael Gieding: Kombinatorik und Geometrie. Diagonaleigenschaften des Vierecks mit dem Heidelberger Winkelkreuz entdecken. IN: PM, Praxis der Mathematik in der Schule, Jahrgang 57, Nr. 61 (Februar 2015), S. 12–18

Das Heidelberger Winkelkreuz ist ursprünglich aus dem Geobrett entstanden, hat gegenüber diesem jedoch eine stärkere dynamische Komponente, mit ihm lassen sich u. a. Vierecke spannen. Der Artikel zeigt, wie Schülerinnen und Schüler mit dem Heidelberger Winkelkreuz die Diagonaleigenschaften verschiedener Viereckarten entdecken können und dabei kombinatorische und geometrische Überlegungen verbinden und ihr Wissen zu den Viereckarten vertiefen können. (Autorreferat)

Hans-Wolfgang Henn: Von Daten zur Funktion. Passende Modelle finden-durch Linearisierung. In: mathematiklehren 187 (Dezember 2014), S. 12–16

Die Anpassung von Funktionen an gemessene Daten-Tupel ist eine wichtige Aufgabe in vielen Wissenschaften. Die theoretische Grundlage dafür ist die von Gauß entwickelte Methode der kleinsten Quadrate. Diese wird oft nur syntaktisch unter Verwendung des sogenannten FIT-Befehls am Computer abgearbeitet. Die in diesem Beitrag behandelte Methode der Linearisierung zeigt einfache dahinter steckende Ideen der Kurvenanpassung, die durchaus gut mit Bleistift und Papier erkundet werden können. (Fazit des Autors)

Nils Hesse: Spielend gewinnen. Wiesbaden: Springer Fachmedien, 2015 (Springer Spektrum), ISBN: 978-3-658-04440-4

Das Buch fasst konkret und verständlich die wichtigsten Gewinnstrategien für die 50 bekanntesten Karten-, Brett-, Würfel-, Karten- und Gewinnspiele zusammen, die sofort angewandt werden können. Griffige Faustformeln und die wichtigsten mathematischen Berechnungen zeigen: Der Weg zum Gewinn führt nicht über Zufall und Glück, sondern über Logik und Strategie. Der Inhalt: 1. Klassische Brettspiele: z. B. Schach, Dame, Backgammon, Scrabble, 2. Kinder- und Familienspiele: z. B. Malefiz, Mensch ärgere Dich nicht, Vier gewinnt, Hase und Igel, 3. Gesellschaftspiele: z. B. Siedler von Catan, Monopoly, Carcassonne, Risiko, Scotland Yard, 4. Kartenspiele: z. B. Doppelkopf, Skat, Poker, Mau-Mau, 5. Würfel-, Tipp-, Wett- und Gewinnspiele: z. B. Kniffel, Fußball-Tipprunden, Lotto, Roulette.

Andreas Kaufmann: Vernetzungen und Kernideen: Ein Minimalprogramm für die Stochastik in der Sekundarstufe II. In: mathematiklehren Nr. 188 (Februar 2015), S. 42–45

Anhand des Vorgehens nach Kernideen, eines durchgängigen Beispiels und des Häufigkeitskonzeptes werden die zentralen Inhalte der Stochastik eingeführt und miteinander vernetzt. Es werden auch Wege zum Erweitern und Verfeinern gezeigt. Die Einheit spannt den Bogen von der Berechnung einer Wahrscheinlichkeit in einem mehrstufigen Zufallsexperiment, Erwartungswert einer Zufallsgröße, faire Wetten, bedingte Wahrscheinlichkeit, Binomialverteilung bis zum Testen einer Hypothese.

Hubert Langlotz; Heinz Laakmann: Von der beschreibenden zur beurteilenden Statistik. In: PM, Praxis der Mathematik, Jahrgang 56 (Dezember 2014) Heft 60, S. 10–13

Sachgerechtes Sammeln, Darstellen und Auswerten von Daten, dies sind wichtige Kompetenzen, die im Stochastikunterricht gelernt werden sollen. Mit angemessenen Darstellungen können schon viele Fragen beantwortet werden, jedoch nicht die Frage, ob die Daten nicht auch zufällig entstanden sein könnten. Dies ist das Gebiet der beurteilenden Statistik. Das Testen von Hypothesen stellt allerdings nicht nur aus Schülersicht oft eines der großen Probleme im Mathematikunterricht der Sekundarstufe II dar. Simulationen helfen, Verständnis dafür zu erzeugen, ob ein Datensatz als zufällig eingestuft werden kann oder nicht. Dabei wird der Mehrwert eines Rechnereinsatzes deutlich. (Autorenreferat)

Andreas Quatember: Statistischer Unsinn: Wenn Medien an der Prozenthürde scheitern. Berlin: Springer, 2015

Vier von zehn oder jeder Vierte …

Ein Blick in eine beliebige Tageszeitung genügt: Statistiken sind ohne Zweifel ein wesentlicher Bestandteil unserer Informationsgesellschaft. Dennoch ist das Image des Faches Statistik denkbar schlecht. Die Diskrepanz zwischen offenkundiger Bedeutung und schlechtem Ruf beruht zum Teil auf dem fundamentalen Irrtum, die Qualität der statistischen Methoden mit der Qualität ihrer Anwendung zu verwechseln. Denn ob aus Unachtsamkeit, Unverständnis oder Unvermögen: In den Medien wird mit Statistiken allzu oft Des-Information statt Information betrieben. Dieses Buch lädt die Leser zu einer kritischen und amüsanten Irrfahrt durch falsche Schlagzeilen und unsinnige Interpretationen statistischer Ergebnisse in Tageszeitungen oder Zeitschriften ein. Staunen Sie darüber, dass ein Viertel aller Studierenden alkoholabhängig ist, dass Männer ihren Rasierern treuer sind als ihren Partnerinnen, dass höherer Schokoladenkonsum mehr Nobelpreisträger erzeugt – und warum das alles blanker Unsinn ist. (Klappentext des Verlags)

Waltraud Schillig: Qualifizierte Frauen sind da! Prozentangaben und bedingte Wahrscheinlichkeiten. In Mathematik 5–10, 29/2014, S. 40–41

Das Thema Frauenanteile in Führungsebenen wird an Hand eines Zeitungsartikels diskutiert. Benötigte Kenntnisse: Auswerten von Diagrammen, Vierfeldertafel, Baumdiagramme, bedingte Wahrscheinlichkeiten.

Ute Sproesser: Heute einmal anders herum! Schülerinnen und Schüler interpretieren Baumdiagramme. In: Mathematik 5–10, Heft 27 (2. Quartal 2014), S. 30–31

Ein Baumdiagramm mit eingezeichneten Wahrscheinlichkeiten wird vorgegeben. Schüler der Jahrgangsklassen 9–10 sollen Beispiele für Sachsituationen als Interpretationsvorschläge für das Baumdiagramm geben.

Heinz Klaus Strick: Welche Erfolgswahrscheinlichkeit liegt dem Zufallsversuch zugrunde? Mithilfe von Simulationen in Fragestellungen der Beurteilenden Statistik einsteigen. In: PM, Praxis der Mathematik, Jahrgang 56 (Dezember 2014) Heft 60, S. 10–13

Im Artikel wird erläutert, wie man mithilfe von Simulationen durch Zufallszahlen erste Erfahrungen mit dem Streuverhalten von Bernoulli-Versuchen sammeln kann. Dies geschieht zunächst durch eine Untersuchung von Boxplots, dann durch Untersuchungen weiterer Perzentilen, welche die Möglichkeit eröffnen, auch ohne vorherige Behandlung der Standardabweichung und der Sigma-Regeln in die Grundfragen der Beurteilenden Statistik einzusteigen. (Autorenreferat)

Kim-Alessandro Weber; Gunnar Friege; Rüdiger Scholz: Ich föhne mir echte Zufallszahlen. In: MNU, Der mathematische und naturwissenschaftliche Unterricht, Jahrgang 68 (Januar 2015) 1, S. 9–11

Die Erzeugung von Zufallszahlen mittels manueller Zufallsgeneratoren (Würfel) ist zeitintensiv und wird kaum betrieben. Die Verwendung der Zufallsfunktion des Computers oder Taschenrechners ist hier eine Lösung, diese Zufallszahlen sind jedoch „Pseudo“, weil zur Generierung ein Algorithmus benutzt wird. In diesem Beitrag wird eine Alternative aufgezeigt, die aus drei Komponenten besteht, die an jeder Schule zur Verfügung gestellt werden können: Mit der Hilfe eines Föhns, Mikrophons und Computers wird ein Generator echter Zufallszahlen realisiert.