Manfred Borovcnik, Klagenfurt

Kritik und Gegenkritik einer ‚unlösbaren' zentralen Abituraufgabe – Wird die Zentralmatura zu einem juristischen Text?

In Nordrhein-Westfalen wurde die ‚Nowitzki-Aufgabe' in ‚unlösbarer' Form gestellt. Das hat zu einer erbitterten Diskussion um das Zentralabitur geführt. Natürlich ist es peinlich, wenn solch ein Fehler unterläuft. Allerdings zerfällt die geäußerte Kritik bei näherer Inspektion. Wie genau man auch formuliert, fast jede Aufgabe kann man mehrdeutig auffassen. Das hat für zentral gestellte Aufgaben weit reichende Folgen, mehr als für lokale. In der Diskussion wurde aber kaum beachtet, dass die verwendete Modellierung nicht angemessen ist und damit die gestellten Fragen kaum sinnvoll beantworten lässt.

1. Vorbemerkungen

Die angesprochene Aufgabe war Teil des zentralen Abiturs in Nordrhein-Westfalen; sie stellt keinesfalls einen Einzelfall dar – alljährlich wiederholt sich dies Spiel von Aufgabenstellung und Kritik. Die Auseinandersetzung mit dieser Aufgabe war jedoch besonders heftig. Vorgestellt wird eine Kritik von Davies, die von einer Gruppe von Statistikern gemeinsam getragen wurde. Diese Kritik wird analysiert. Als Vorausblick auf die Ergebnisse sei schon einmal festgehalten:

- Die Kritik von Davies zerfällt bei kritischer Betrachtung.
- Es ist schwierig, zentral Aufgaben zu stellen.
- Zentral gestellte Aufgaben müssen notgedrungen auf Basiskompetenzen fokussieren.
- Die Textgestaltung ufert zu einer juristischen Aufgabe aus.

Weder Aufgabensteller noch die folgende Diskussion stoßen sich daran, dass die gewählte Modellierung die anstehenden Fragen – auch wenn die fehlenden Angaben ergänzt werden – kaum sinnvoll beantworten lässt. In dieser Diskussion wird vehement – wie bei Davies (2009) – vertreten: „Wenn die Modellierung feststeht, wird die Aufgabe innerhalb des Modells weiter bearbeitet. […] angenommen, […] Modell stimmt. Es ist nicht die Aufgabe des Abiturienten, die Angemessenheit der Modellierung zu hinterfragen." (S. 2) Das hat zur Folge: Das Modell wird eindeutig vorgegeben, seine Adäquatheit wird einfach unterstellt. Gerade die interessanten Modellierungsaspekte, ein Hinterfragen, warum und ob die Ergebnisse denn auch relevant sein könnten, werden peinlich vermieden. Übrig bleibt schematisches ‚Anwenden' statistischer Methoden, die auf Basiskompetenzen reduziert werden.

Wenn Fragen der Angemessenheit der verwendeten Modelle im Abitur so strikt ausgeklammert werden, bedeutet das zwar noch nicht, dass solche Fragen im Unterricht wegfallen müssen; es kann aber vermutet werden, dass sich dann auch im Unterricht eine Tendenz einstellt, Anwendungen auf ‚Anwendungen' zu reduzieren. In diesem Aufsatz soll die Kritik an der Aufgabenstellung und an der Kritik genützt werden, um aus dem Blickwinkel der Modellierung auf die Stochastik zu schauen. Welche Qualifikationen von Seiten der Schüler sind gefragt, um stochastische Begriffe anwenden zu können? Welche Sichtweisen von Mathematik sind zu entwickeln, damit man mit den Ergebnissen sinnvoll umgehen kann? Gleichzeitig wird klar werden, dass Aufgaben aus einem Kontext implizit Mehrdeutigkeiten enthalten. Das ist im Hinblick auf den Modellbildungsprozess durchaus wünschenswert, kommt aber mit den Zielsetzungen einer zentral gestellten Matura in Konflikt, weil eine zentrale Aufgabenstellung unmissverständlich formuliert sein muss und zwar sowohl hinsichtlich der Fragestellung als auch hinsichtlich der Modellierung. Wenn über das zu verwendende Modell zweifelsfrei Klarheit

bestehen muss, dann wird die Lösung aus der Sicht von Modellbildung trivial, weil sie sich auf das Ausrechnen der Lösung innerhalb eines festen, vorgegebenen Modells beschränken muss.

Im zweiten Abschnitt wird die Aufgabe vorgestellt und ihre Modellierung mit Bernoulli-Ketten (unabhängige 0-1-Experimente mit durchgängig derselben Erfolgschance) erläutert. Dabei tritt die Problematik auf, dass man gelegentlich von wahren Werten einer (allenfalls unbekannten) Wahrscheinlichkeit spricht. Eine Schätzung von ‚wahren' Wahrscheinlichkeiten, die sich zudem je nach modellierten Einflussgrößen unterscheiden, erfordert, dass der Prozess der Entstehung der Daten tatsächlich einer Bernoulli-Kette folgt.

In der Praxis wird statt einer Schätzung oft ein fiktiver Wert für diese Wahrscheinlichkeit *unterstellt*. Das entspringt nicht nur der Willkür sondern entspricht der Umsetzung von Fragen; etwa, ist der Spieler noch so stark wie in der abgelaufenen Saison? Entsprechend dem Einbringen solcher Information über die Stärke des Spielers werden die einfachen probabilistischen Aufgaben gelöst. Insbesondere wird auch der Frage nachgegangen, wie man die Ungenauigkeit der Schätzung der Spielstärke (als Wahrscheinlichkeit aufgefasst) evaluieren kann.

Im dritten Abschnitt wird trotz der schon eingebrachten Kritik an den geforderten Voraussetzungen der ‚unlösbare' Teil der Aufgabe mit dem Bernoulli-Modell besprochen. Dabei werden auch die innerhalb der öffentlichen Kritik vorgebrachten Versuche, die Aufgabe abzuändern und damit zu lösen, kritisch gewürdigt. Danach wird an eine fundamentale, aber in dieser Form wenig bewusst erfasste Eigenschaft von Bernoulli-Ketten erinnert: Jedes Spiel markiert einen neuen Anfang! Mit einem Schlag wird die Aufgabe – ganz einfach – lösbar. In einer der vorgebrachten Varianten der Aufgabe wird das entstehende kombinatorische Problem als ziemlich schwierig angeprangert. Das hat den Autor dazu veranlasst, daran zu erinnern, dass die Kombinatorik eine Kunst guten Zählens ist. Zählt man schlecht, wird die Aufgabe immer schwierig sein.

Im vierten Abschnitt wird der statistischen Frage nachgegangen, ob Nowitzki auswärts schlechter ist als daheim. Im Versuch, dem Ministerium als Aufgabensteller Fehler vorzuwerfen, fordern die Kritiker eine strikte Trennung von ‚wahren' Werten und Schätzwerten, verstricken sich dann aber in einer völlig unstatistischen Betrachtungsweise, wie man bei einem statistischen Test vorgeht. Sie kommen mit ihrem Versuch in eine Sackgasse, weil dabei häufig auch gar keine *wahren* Werte für einen Parameter zum Test anstehen, sondern *fiktive* – die etwa Fragestellungen aus dem Kontext entspringen können, oder weil wahre Werte sowieso immer von einem verwendeten Modell abhängen.

Neben dem Binomial-Modell werden andere, in der Praxis übliche Verfahren zur Behandlung der Kernfrage vorgeschlagen. Der Fisher'sche exakte Test benötigt nur Kenntnisse über die hypergeometrische Verteilung. Diese Verteilung ist zwar nicht im Kanon der Schule, unterscheidet sich aber im Schwierigkeitsgrad nicht von der Binomialverteilung, mit der sie über das Ziehen von Kugeln aus einer Urne eng verwandt ist. Darüber hinaus wird das Nowitzki-Problem in den Prozess der empirischen Erkenntnisgewinnung eingebettet. Während im probabilistischen Teil die Voraussetzungen einer Bernoulli-Kette wenig überzeugend wirken, kommt – wie die Überlegungen zeigen werden – einem heuristischen Argument der Homogenisierung und Ausgleichung von Verletzungen der Annahmen bei globalerer statistischer Betrachtungsweise doch einige Kraft zu.

Im fünften Abschnitt werden die Modellierungsschritte nocheinmal im Hinblick auf die erforderlichen Kompetenzen beleuchtet. In einer vollständig zentral gestellten schriftlichen Matura scheinen diese keinen Platz zu haben. Man kann das an dieser Stelle schon bedauern; man kann aber auch die Frage aufwerfen, ob sich das noch lohnt zu unterrichten, was dann von Stochastik in der Schule übrig bleibt.

2. Die Aufgabe und erste Modellierungsschritte

Zuerst wird die Nowitzki-Aufgabe aus dem Sport wiedergegeben, dann wird ausführlich seine Modellierung erläutert. Erst durch weitere Informationen aus dem Kontext werden die noch offenen Parameter festgelegt (oder geschätzt) und damit das Modell eindeutig fixiert. Diese Parameter als ‚wahre' Werte anzusprechen, geht an ihrem Charakter – ihrer Abhängigkeit von der gewählten Modellierung – vorbei. Je nachdem, ob man weitere Einflussgrößen berücksichtigt oder nicht, führt das zu anderen wahren Werten. Fehlt ein Wissen über solche Einflussgrößen oder fehlen Daten darüber, so entzieht sich diese Abhängigkeit der wahren Werte einer Analyse. Die Diktion von *wahren* Werten täuscht die Illusion einer wesensmäßigen Bindung von Modell und Realität vor und verengt den Interpretationsspielraum unnötig.

Bei vielen Aufgaben aus dem Bereich der Wahrscheinlichkeitsrechnung wird beklagt, dass es lediglich um Routinetechniken geht; ein Kontext spielt allenfalls die Rolle einer leeren Hülle (siehe Eichler und Vogel 2010): Berechne, mit Tabellen oder mit Taschenrechner etc. die Wahrscheinlichkeit, ‚höchstens 8 Erfolge bei 10 Versuchen zu bekommen'. Dabei wird zudem noch klar gestellt, dass es um die Binomialverteilung geht und der Parameter p so und so groß ist. Damit ist keine Fragestellung verbunden, das Modell ist fest vorgegeben, es bleibt offen, warum man eine solche Wahrscheinlichkeit ausrechnen soll und was man – hat man sie berechnet – nun besser kann. So gesehen ist es durchaus begrüßenswert, wenn Lehrbücher oder gar zentral gestellte Maturaprüfungen die Aufgaben in einen sinnvollen Kontext zu kleiden und Fragestellungen zu entwickeln versuchen.

Dabei hat sich Sport als motivierender Kontext erwiesen. Dass man aber gerade dabei – nimmt man Modellierung ernst – in große Schwierigkeiten geraten kann, zeigt das Beispiel aus dem deutschen Bundesland. Die grundsätzlichen Voraussetzungen einer Bernoulli-Kette sind im Sport unangemessen.

Die Aufgabenstellung aus dem Kontext des Sports – Nowitzkis Spielstärke

„Der deutsche Basketball-Profi Dirk Nowitzki spielt in der amerikanischen Profiliga NBA [...] In der Saison 2006/7 erzielte er bei Freiwürfen eine Trefferquote von 90,4%.

a) Berechnen Sie die Wahrscheinlichkeit, dass er
   (1) genau 8 Treffer bei 10 Versuchen erzielt,
   (2) höchstens 8 Treffer bei 10 Versuchen erzielt,
   (3) höchstens viermal nacheinander bei Freiversuchen erfolgreich ist.

b) Bei Heimspielen hatte er eine Freiwurfbilanz von 267 Treffern bei 288 Versuchen, bei Auswärtsspielen lag die Quote bei 231:263. Ein Sportreporter berichtet, dass Nowitzki auswärts eine deutlich schwächere Freiwurfquote habe. Untersuchen Sie auf einem Signifikanzniveau von 5%, ob die Trefferanzahl bei Auswärtsspielen
   (1) signifikant unter dem Erwartungswert für Heim- und Auswärtsspiele liegt,
   (2) signifikant unter dem Erwartungswert für Heimspiele liegt."

(Schulministerium NRW, o.D.)

Tabelle 1: Scores der Saison 2006/7.

Spiele	Treffer	Versuche	Erfolgswahrscheinlichkeit – faktisch bekannt oder geschätzt
Heim	$T_H = 267$	$n_H = 288$	$p_H =$ bzw. $\approx \frac{267}{288} = 0{,}927$
Auswärts	$T_A = 231$	$n_A = 263$	$p_A =$ bzw. $\approx \frac{231}{263} = 0{,}878$
Alle	$T = 498$	$n = 551$	$p_g =$ bzw. $\approx \frac{498}{551} = 0{,}904$

Modellierung 1 – Bernoulli-Ketten

Wahrscheinlichkeit hängt eng mit der Deutung als relative Häufigkeit zusammen; der frühe Beweis dafür von Bernoulli 1714 hat aber schon gezeigt, dass diese Deutung von *Voraussetzungen* abhängt. Nicht immer hat es einen Sinn, Häufigkeiten als Schätzung einer Wahrscheinlichkeit zu interpretieren. Wahrscheinlichkeiten, die zahlenmäßig bekannt sind, hängen von einem *Modell* (etwa einer Gleichverteilung) ab. In diesem Abschnitt werden damit zusammenhängende Fragen erörtert.

Mit der Binomialverteilung modelliert man Vorgänge,
- bei denen eine feste Anzahl n von Versuchen durchgeführt wird;
- bei jedem Versuch gibt es nur zwei Möglichkeiten (Erfolg, Misserfolg).

Will man das Modell auf den Kontext anwenden, muss man davon ausgehen, dass
- die Erfolgswahrscheinlichkeit p für alle Versuche dieselbe ist, und
- dass sich die einzelnen Versuche nicht gegenseitig ‚beeinflussen'.

Der Prozess der Ereignisse, welcher die einzelnen Versuche beschreibt, wird auch als Bernoulli-Kette bezeichnet. Die einzelnen Zufallsvariablen X_i nehmen nur die Werte 0 (‚Misserfolg') bzw. 1 (‚Erfolg') an und sind unabhängig voneinander; mathematisch schreibt man dafür

$$X_i \overset{iid}{\sim} B(1, p) .$$

Dabei steht das Akronym ‚iid' für ‚independent, identically distributed' und die Tilde für ‚ist verteilt nach'. Der Parameter p – die Erfolgswahrscheinlichkeit – ist mal bekannt, mal muss er aus Daten geschätzt werden, mal wird er fiktiv (je nach Fragestellung) unterstellt. Um die Argumente leichter zuordnen zu können, wird im Folgenden der Kern der Kritik der Aufgabe aus Davies (2009) immer in der linken Spalte wiedergegeben und in der rechten kommentiert.

Wahre Werte und Schätzwerte

„Wirft man eine 1€-Münze, beträgt die Wahrscheinlichkeit für einen Zahlwurf ½. […] normalen Würfel, […] bestimmte [Zahl] ⅙. […] Diese Wahrscheinlichkeiten werden durch Symmetrieüberlegungen ermittelt." (S. 2)

„Bei einer normalen Münze scheint [½] aus Symmetriegründen plausibel, denn es gibt im Normalfall keinen Grund, die eine oder die andere Seite […] vorzuziehen. Ähnliches gilt für einen Würfel oder […]. Um ganz eindeutig zu sein, […] "fairen" oder "unverfälschten" Münze sprechen." (S. 3)

„[…], wo es keine Symmetrieargumente gibt, z.B. die Erfolgswahrscheinlichkeit bei Freiwürfen im Basketball, greift man auf empirische Ergebnisse zurück und die Wahrscheinlichkeit wird geschätzt." (S. 2)

Davies fordert ein, „wahre Wahrscheinlichkeiten" und „Schätzwerte" sauber zu trennen.

Wahre oder fiktive Werte

Kann man denn Werte für eine Wahrscheinlichkeit ohne Bezug auf ein Modell angeben?

Der Kontext der Glücksspiele diente zur ersten Definition von Wahrscheinlichkeit bei Laplace (1812). Wahrscheinlichkeit schien damit als physikalische Eigenschaft eines Würfels eindeutig festgelegt. Bei jedem echten Würfel würde man eine Abweichung davon feststellen. Jede Wahrscheinlichkeitsaussage ist an ein Modell gebunden. Es bleibt ein Modell, selbst wenn es gut passt.

Kann man eine Erfolgswahrscheinlichkeit im Sport sinnvoll unterstellen? Und kann man diese aus relativen Häufigkeiten dann auch schätzen?

Eine Interpretation von Wahrscheinlichkeit als relativer Häufigkeit bedingt, dass die Daten dazu in einer Bernoulli-Kette entstehen. Im Glücksspielbereich hat man keine Probleme mit derselben Erfolgswahrscheinlichkeit und mit der Unabhängigkeit der Versuche. Aber im Sport?

Die Daten in Tabelle 1 geben einen Überblick über die Treffer in den jeweiligen Spielen in der Saison 2006/7. Kann man die Scores 0,927 (daheim), 0,878 (auswärts) und schließlich 0,904 in allen Spielen als Stärke des Spielers interpretieren? Gibt es eine wahre Stärke des Spielers, oder hängt diese von weiteren Faktoren ab, etwa, wo das Spiel stattfindet. Unter welchen Umständen hat das einen Sinn?

Wird die Saison als Einheit betrachtet, dann beschreiben die Daten die Stärke des Spielers. Die Frage ‚ist er schwächer auswärts als daheim?' kann nur rhetorisch gemeint sein; er *war* es und zwar deutlich: die Erfolgsraten unterscheiden sich um beinahe 5%-Punkte (0,927 – 0,878). Die Frage, ob der Unterschied signifikant ist, ist nicht zulässig. Es ist überhaupt zweifelhaft, ob die Erfolgsraten als Wahrscheinlichkeit interpretierbar sind. Im Abschnitt Modellierung 3 (weiter unten) wird thematisiert, wie man Information über die Heimstärke einbringt. Kennt man diese Heimwahrscheinlichkeit (oder unterstellt man sie, woher auch immer die Information kommt), so wird die Frage nach einer signifikanten Abweichung bei Auswärtsspielen möglich; dieses Thema wird in Abschnitt 4 aufgenommen.

Wenn die Erfolgsraten als Wahrscheinlichkeiten interpretiert werden sollen, dann muss man Nowitzkis Freiwürfe als Exemplar einer Bernoulli-Kette – mit derselben Erfolgswahrscheinlichkeit und der Unabhängigkeit zwischen den Würfen – auffassen. Nur so kann man garantieren, dass die relativen Häufigkeiten als Schätzwert einer unbekannten Erfolgswahrscheinlichkeit gelten können. Aber schon die Erfolgsraten in Heim- und Auswärtsspielen (Tabelle 1) unterscheiden sich erheblich voneinander. Die Erfolgswahrscheinlichkeit mag auch von anderen Größen beeinflusst werden.

Modellierung 2 – Confounder beeinflussen eine Wahrscheinlichkeit

Daten bedürfen immer einer Interpretation. Häufigkeiten müssen nicht aus einem homogenen Prozess der Datengewinnung stammen. Störfaktoren können sie überlagern. Ob man Unterschiede in Daten als Folge ‚natürlicher' Streuung von Zufallsprozessen interpretiert oder sie auf Störgrößen zurückführt, verändert gänzlich die Interpretation vorliegender ‚Fakten'.

Nimmt man an, dass Heim- und Auswärtsspiele gleich sind, könnte man einen gemeinsamen Prozess der Datenerzeugung mit $p = 0{,}904$ als Modell verwenden und sich wundern, wie stark die Daten fluktuieren: die Erfolgsrate in 288 Versuchen daheim beträgt 0,927, in 263 Versuchen auswärts nur 0,878. Beide Werte liegen gerade noch innerhalb der üblichen Marge von Zufallsschwankungen.

Ein weiteres Beispiel soll die Rolle von Confoundern (Störgrößen) illustrieren.

Beispiel: Die Anteile an Mädchengeburten sind in Tabelle 2 angegeben (Daten bis auf die letzte Zeile aus Davies 2009).

Tabelle 2: Mädchenanteil in verschiedenen Spitälern.

Spital	Geburten	Anteil der Mädchen
A	514	0,492
B	358	0,450
C		0,508
'Welt-Statistik'		0,489

Wahre Werte – und Schätzungen davon

Davies stellt fest: „Diese drei Zahlen können nicht alle die Wahrscheinlichkeit für ein Mädchen sein, sonst hinge diese Wahrscheinlichkeit vom Krankenhaus ab. […] Will man ein Mädchen, dann Krankenhaus *C*, will man einen Jungen, dann Krankenhaus *B*. In der mathematischen Statistik unterscheidet man zwischen einer wahren Wahrscheinlichkeit und einer geschätzten Wahrscheinlichkeit. […]

Werte hängen immer vom Modell ab

Unterstellt sei eine Wahrscheinlichkeit für Mädchen weltweit mit 0,489. Dann liegt der Anteil an Mädchen im Spital *B* (bei 358 Geburten) mit 95% Wahrscheinlichkeit zwischen 0,437 und 0,541. Der beobachtete Wert von 0,450 ist recht nahe am unteren Rand dieses Intervalls und wirft Zweifel auf, dass die Daten durch ‚reinen Zufall' entstanden sind, i.e., dass eine Bernoulli-Kette mit $p = 0{,}489$ im Hintergrund steht.

Bei einer Textaufgabe muss die Modellierung […] angegeben werden, oder sie muss aus der Beschreibung der Situation eindeutig hervorgehen.

[…] Modellierung feststeht, wird die Aufgabe innerhalb des Modells […] es wird angenommen, dass das Modell stimmt. Es ist nicht die Aufgabe des Abiturienten, die Angemessenheit der Modellierung zu hinterfragen." (S. 2)

Es mag besser sein, die Situation durch drei *verschiedene* Bernoulli-Ketten zu modellieren. Solche Phänomene treten in der Praxis häufig auf.

Man wird die Unterschiede durch Bezug auf *Kovariate* zu erklären suchen: Eine Spekulation im Kontext betreffend des Orts des Spitals sei angeführt: Spital *A* in Deutschland, *C* in der Türkei und *B* in China.

Das Beispiel erläutert indirekt den Kern des Problems: So lange man den datenerzeugenden Prozess nicht gut kennt und auch keine Kenntnis hat über andere Merkmale, welche das Zielmerkmal (hier Geschlecht des Neugeborenen) beeinflussen, bleiben solche Kovariaten versteckt. Während sie die Ergebnisse beeinflussen, kennt man sie nicht; bekommt man eine Ahnung, dass solche Vorgänge im Hintergrund wirken, kann man sie nicht analysieren, weil man keine Daten dazu hat. In solchen Fällen nennt man diese Merkmale Störgrößen oder *Confounder*. Potentielle Confounder muss man am Beginn einer Studie – also bevor man sich die Daten beschafft – systemisch abklären; sonst hat man später keine Chance, ihren Einfluss zu analysieren, weil Daten dazu einfach fehlen.

Für die Wahrscheinlichkeitsfragen bei der Nowitzki-Aufgabe hat das zur Folge, dass es besser wäre nach Confoundern zu suchen (ob er gut in Form ist, ob er Streit im Team oder mit seiner Ehefrau hatte etc.), um danach eine Wahrscheinlichkeit zu berechnen, ob er mindestens 8mal von 10 Versuchen trifft, anstatt das Problem mit einer Bernoulli-Kette der Länge 10 mit der Stärke des Spielers für die gesamte Saison zu modellieren. Solch ein Zugang kann zwar nicht in einer zentral gestellten Matura geprüft werden, sollte aber sehr wohl Gegenstand des Klassengesprächs zum Thema sein.

Modellierung 3 – Information über die Spielstärke *p*

In einer primitiv frequentistischen Deutung von Wahrscheinlichkeit gibt es nur die Möglichkeit, die Spielstärke durch Daten, die den Voraussetzungen einer Bernoulli-Kette genügen, zu schätzen. In der Statistik wird jedoch oft ein Wert für den Parameter p unterstellt, und zwar aus ‚langjähriger Erfahrung' oder aus Spekulationen. Dann will man vielleicht eine Frage klären: Kann diese Vermutung vom Tisch gewischt werden oder muss man sie in die weiteren Überlegungen einbeziehen? In diesem Abschnitt wird diskutiert, wie man Information über einen Parameter ins Modell einbringen kann.

Aus der Perspektive der Modellierung betrifft die erste Entscheidung die *Familie* von Verteilungen; hier wurde die Bernoulli-Kette genommen und das hat zur Folge, dass die Zahl der Erfolge bei n Versuchen einer Binomialverteilung $B(n, p)$ folgt; allerdings bleibt offen, welchen Wert man für diesen Parameter p einzusetzen hat. Dieser Prozess ist für alle Modellierungen ähnlich, wenngleich sich im Kontext der Aufgabe auch eine Besonderheit ergibt, nämlich dass die Grundgesamtheit auch als endlich angesehen werden kann. Statt alle Spiele einer Saison als Realisierung einer Stichprobe einer oder mehrerer Bernoulli-Ketten aufzufassen, kann man auch alle (endlich vielen) Spiele als gegeben betrachten und fragen: Ist es ungewöhnlich, bei den Auswärtsspielen so wenige Treffer vorzufinden? Wir besprechen drei Fälle, wie man Information über unbekannte Parameter einbringen kann und ziehen einige Konsequenzen daraus für die Problematik von wahren Werten von Parametern.

i. p ist bekannt,
ii. p wird aus den Daten geschätzt,
iii. p wird aus ähnlichen Situationen übertragen oder – gemäß einer Hypothese – fiktiv unterstellt.

i. In Glücksspielen werden Symmetrieargumente herangezogen, sodass relativ zu einem Gleichverteilungsmodell der Wert als bekannt unterstellt wird. In anderen Situationen wird man relative Häufigkeiten als Basis heranziehen und feststellen, dass langjährige Erfahrung den Wert festlegen lässt. Dabei unterstellt man eine ‚ceteris paribus'-Bedingung, wonach alles beim selben geblieben ist. Ansonsten hat eine Übertragung der Häufigkeiten aus der Vergangenheit keinen Sinn.

ii. Man schätzt p aus so aktuell wie möglichen Daten, das ist die gegenwärtige Saison. Folgen die Daten einer Bernoulli-Kette X_i, so schätzt man den Parameter p durch

$$\hat{p} = \overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} = 0{,}904 .$$

Berücksichtigt man, dass die Daten mit einer zufälligen Fluktuation behaftet sind, so kann man aus ihnen ein (1-α)-Konfidenzintervall berechnen, das sich unter Normalapproximation so angeben lässt:

$$\hat{p} \mp z_{1-\alpha/2} \frac{\sqrt{\hat{p} \cdot (1-\hat{p})}}{\sqrt{n}} .$$

Natürlich hat das wenig Sinn, wenn die Voraussetzungen einer Bernoulli-Kette grob verletzt sind.

iii. Eine Hypothese über die Erfolgswahrscheinlichkeit könnte durch Werte aus der Vergangenheit wie in i. unterstellt (und nicht wie in ii. geschätzt) werden. Man könnte sich auf den Standpunkt stellen, dass die gegenwärtige Saison vollständig und daher die Erfolgsrate mit 0,904 bekannt ist. Man modelliert die Daten, als ob sie aus einer entsprechenden Bernoulli-Kette stammen und rechnet in diesem Modell. Die Wahrscheinlichkeit ist hier als faktisch bekannt unterstellt. Man erhält numerisch denselben Wert wie in ii., verbindet damit aber eine ganz andere Vorstellung. Insbesondere ist das keine Schätzung und es hat keinen Sinn, ein Konfidenzintervall wie in ii. zu berechnen. Man könnte auch (im Sinne einer Qualitätskontrolle) prüfen, ob Nowitzki einen Wert von 0,95 einhält – anderenfalls man ihm die Gage kürzen würde.

Wahre Werte und Schätzwerte unterscheiden

„In der mathematischen Statistik unterscheidet man zwischen einer wahren Wahrscheinlichkeit und einer geschätzten Wahrscheinlichkeit. Diese begriffliche Unterscheidung ist von grundlegender Bedeutung und unerlässlich.

In der Abituraufgabe werden die beiden Begriffe vermischt: Manchmal ist die Quote die wahre Wahrscheinlichkeit, manchmal ist sie ein Schätzwert.

Somit ist die Aufgabe schlecht gestellt und die Lösung des Ministeriums falsch. […]

In der letzteren Situation muss klar zwischen dem wahren Wert p und einem Schätzwert $\hat{p}$ unterschieden werden. […]

Die klare Unterscheidung zwischen einem wahren Parameterwert und einem Schätzwert hierfür ist fundamental: die ganze schließende Statistik basiert darauf." (S. 2)

Alle Werte sind relativ zu Modellen

Was ein wahrer Wert ist, hängt vom Modell (oder vom Messverfahren) ab. Speziell in der Teststatistik *unterstellt* man einen Wert aufgrund einer Fragestellung und prüft dann, ob er zutrifft oder ob man ihn ablehnen kann. Man kann sich eigentlich fast nie darauf verlassen, dass der unterstellte Wert der wahre Wert ist.

Es geht also mehr darum, wie man Wissen über p formuliert, und wie man es überprüft bzw. testet. Es gibt, wie besprochen, viele Wege, sich entsprechende Information über p zu verschaffen. Daten aus einer Stichprobe und die Schätzung sind *ein* Weg. Allerdings: Eine solche Schätzung setzt voraus, dass man die Daten aus einer Bernoulli-Kette erhält; das trifft hier keineswegs zu.

Von fiktiven Werten statt von wahren Werten zu sprechen, ist daher viel zielführender. Klare Unterscheidung ja, wahrer Parameterwert nein – denn dieser hängt vom verwendeten Modell ab.

Die triviale Lösung, wenn alles feststeht

Gemäß den Möglichkeiten, Information über den unbekannten Erfolgsparameter einzubringen, wird der probabilistische Teil der Aufgabe jetzt mit verschiedenen Modellen gelöst. Die Evaluation der Genauigkeit der Schätzung des Parameters p überrascht: Die Lösungen für den günstigsten und den ungünstigsten Fall unterscheiden sich gewaltig, obwohl die Schätzung auf ziemlich vielen Daten (mehr als 500) basiert. Für die Summe der Erfolge (Treffer) $T_n := X_1 + X_2 + \ldots + X_n$ hat man eine Binomialverteilung, d. h., $T_n \sim B(n, p)$. Wir untersuchen insgesamt drei Modelle:

Modell 1: In der letzten Saison hatte Nowitzki eine Erfolgsrate von 0,910; wenn alles gleich geblieben ist, fixiert man das Modell als $T_n \sim B(n, 0{,}910)$.

Modell 2: In der Saison 2006/7 hatte Nowitzki eine Erfolgsrate von 0,904; wir schätzen den Parameter damit und fixieren das Modell als $T_n \sim B(n, \hat{p} = 0{,}904)$.

Modell 2 Variante: Wir tragen den Ungenauigkeiten des Schätzwerts Rechnung und bestimmen ein 95%-Konfidenzintervall; aus $1-\alpha = 0{,}95$ berechnen wir zuerst $1-\alpha/2 = 0{,}975$ sowie $z_{1-\alpha/2} = 1{,}96$ und erhalten: $[p_U, p_O] = [0{,}8792, 0{,}9284]$. Wir werden gelegentlich p_U und p_O als worst und best case ansprechen und das Modell des worst case lautet dann: $T_n \sim B(n, p_U = 0{,}879)$.

Modell 3: Aus der abgeschlossenen Saison wird die Spielstärke gleich gesetzt mit der festgestellten Erfolgsrate: $p := 0{,}904$ und das Modell ist – trotz gänzlich anderer Interpretation – gleich lautend mit Modell 2: $T_n \sim B(n, p := 0{,}904)$.

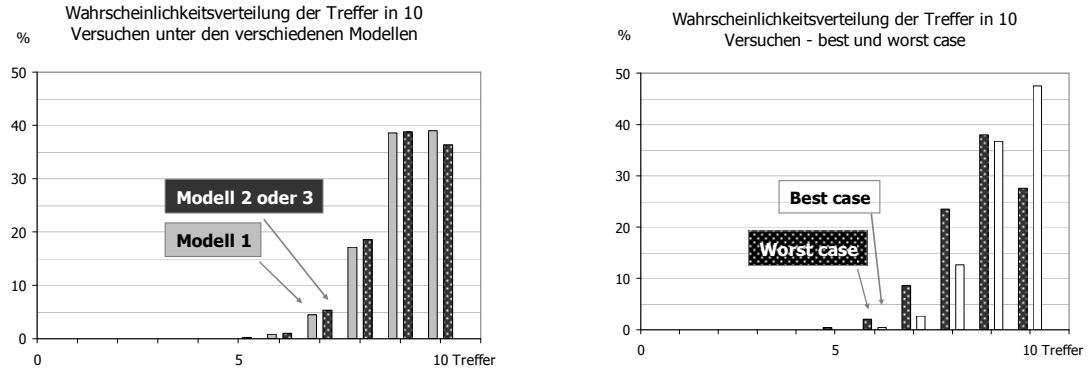

Abb. 1: Wahrscheinlichkeitsverteilung für die Zahl der Erfolge unter den verschiedenen Modellen.

Tabelle 3: Lösungen der ersten beiden Wahrscheinlichkeitsaufgaben.

	Modell 1 $p = 0{,}910$	Modell 2 (3) $\hat{p} = 0{,}904$	Modell 2 (Variante) Worst case p_U	Best case p_O
$P(T_{10} = 8)$	0,1714	0,1854	0,2345	0,1273
$P(T_{10} \le 8)$	0,2254	0,2492	0,3449	0,1573

Während Modell 1 und 2(3) sich im Ergebnis kaum unterscheiden, überrascht die Bandbreite der Ergebnisse, die sich aus der Berücksichtigung der Unsicherheit der Schätzung ergibt. Obwohl wegen der vielen Daten die Schätzung für p im Konfidenzintervall nur einen Spielraum von 5%-Punkten hat, ist die gesuchte Wahrscheinlichkeit, höchstens 8 Erfolge zu haben, nur zwischen 0,1573 und 0,3449 (bei $1-\alpha = 0{,}95$) festzumachen. Solche Sensitivitätsanalysen wären immer angebracht, wenn man Parameter aus Daten schätzt, um das verwendete Modell zu fixieren. Die Praxis bietet leider ein anderes Bild: Mehrfachanalysen werden kaum durchgeführt – nicht zuletzt, weil die Erwartungshaltung nach *einem* Ergebnis stark ausgeprägt ist und man mit parallelen Ergebnissen wenig anfangen kann.

Das Märchen von besseren Schätzwerten durch größere Datenmengen

Will man die Präzision einer Schätzung verbessern, so fordert man einfach mehr Daten ein. Weil aber in der Praxis dadurch die Voraussetzungen an die Entstehung von Daten eher verletzt werden, geht dadurch häufig die Qualität der Daten und damit der Schätzung in *unkontrollierter* Weise verloren. Im Kontext des Sports kann man die Zahl der Daten nicht beliebig erhöhen; mischt man dann Daten aus unterschiedlichen Saisonen, so ergeben sich meist gröbere Abweichungen von den Voraussetzungen der gleichen Trefferwahrscheinlichkeit.

*Präzision nimmt mit 1/Wurzel (*n*) zu*

„Bei Geburten stellt man anhand der Daten ziemlich schnell fest, dass es mehr Jungen als Mädchen gibt. Will man die Wahrscheinlichkeit für eine Mädchengeburt bestimmen, so muss man so viele unverfälschte Daten wie möglich, die über Zeit und Raum homogen sind, bekommen, und daraus die relative Häufigkeit von Mädchengeburten bestimmen.

Der Einfachheit halber nehmen wir an, dass wir die Daten für eine Millionen Geburten haben, wovon 481.345 Mädchen waren.
Man könnte nun p durch die Zahl 0,481345 *schätzen* aber man weiß, dass dies nicht der ”wahre“ Wert von p ist […]

In der Statistik gibt man deswegen ein sogenanntes Konfidenzintervall von plausiblen Werten für p an. […] erhalten wir […] für p das Intervall [0,4804, 0,4823]“ (S. 3)

Mehr Daten – weniger Präzision

Man kann die mit einer bestimmten Stichprobe erreichte Genauigkeit durch ein Konfidenzintervall evaluieren. Es ist verlockend, sich mehr Daten zu besorgen, um die Genauigkeit zu erhöhen. In der Praxis geht aber eine Erhöhung der Zahl oft mit einer Verringerung der Qualität der Daten einher, weil die Voraussetzungen verletzt werden. Es gibt dann kaum Möglichkeiten, die entstandenen Verzerrungen zu korrigieren.

Für Nowitzki müsste man Daten aus mehreren Saisonen zusammenlegen. Schon bei Heim- und Auswärtsspielen scheint klar, dass verschiedene Erfolgswahrscheinlichkeiten ein besseres Modell ergeben. Mehrere Saisonen zu vermengen, hat wenig Sinn. In vielen Fällen ist es aussichtslos, die Genauigkeit einer Schätzung zu verbessern. Man kann lediglich die bestehende Genauigkeit durch worst/best case-Analysen evaluieren.

3. Nowitzkis Chance, höchstens viermal hintereinander zu treffen

Jetzt wird der ‚unlösbare’ Teil der Aufgabe behandelt. Die Kritik an der offiziellen Lösung zeigt gravierende Mängel. Auch die Art der Kritik ist merkwürdig: Es wird übergangen, dass die Modellierung fragwürdig ist; man klinkt sich bei der Unlösbarkeit ein: das darf nicht passieren. Auch auf die Kombinatorik, die in den Reparaturversuchen der Aufgabe als zu schwer eingestuft wurde, wird eingegangen. Die Lösung ist aber letztendlich ganz einfach. Es mag verwundern, dass dies im Tumult der Auseinandersetzung übersehen wurde.

„Ministerium bei der Erstellung von Mathe-Aufgaben im Zentralabitur überfordert? […] für den Leistungskurs Mathematik des diesjährigen Zentralabiturs ist eine Aufgabe über die Freiwürfe des Basketballers Dirk Nowitzki. In der Presse wurde berichtet, dass eine Teilaufgabe nicht lösbar sei, weil die Anzahl der Versuche in der Aufgabe fehlte. Dies ist in der Tat der Fall, und somit sind mehrere Interpretationen der Textaufgabe möglich, die zu verschiedenen Lösungen führen. Die Schulen erhielten vom Ministerium Lösungsskizzen […]. Die Lösung des Ministeriums ist nur dann richtig, wenn die Anzahl der Freiwürfe fünf ist. Aus der Aufgabenstellung gibt es aber keinen Grund, dies anzunehmen.“ (Davies, L., Dette, H., Diepenbrock, F.R., & Krämer, W., 2008)

Unlösbarkeit der dritten Teilaufgabe als öffentlicher Streitpunkt

Die dritte Frage im Wahrscheinlichkeitsteil der Aufgabe wurde in der Öffentlichkeit heftig diskutiert. Ihre Unlösbarkeit wurde als Skandal angeprangert. Auch die genannten Statistiker haben aktiv am Geschehen teilgenommen; sie hatten zwar noch weitere Motive – die Vermengung von ‚wahren Werten' und ‚Schätzwerten' in der Aufgabenstellung –, aber auch sie kritisierten die Aufgabensteller in einem offenen Brief ob der Unlösbarkeit (siehe das Zitat).

Davies unternimmt drei Versuche, die Aufgabenstellung zu reparieren:

1. „Nimmt man an, dass 10 Versuche […]"
2. „[…] $n = 5$ ist"
3. „[…] beim Training wird folgendes Spiel zwischen zwei Spielern [vereinbart]"

1. „Von den $2^{10} = 1024$ möglichen Versuchsreihen muss man die Anzahl […] bestimmen, in denen die Eins höchstens viermal nacheinander vorkommt. […] zuerst die Gesamtanzahl k von Erfolgen festlegt.

Bei $k = 10$ und $k = 9$ […], dass es keine Möglichkeit gibt, die Einsen so zu platzieren, dass sie höchstens viermal nacheinander vorkommen.

Bei $k = 8$ gibt es 15 Möglichkeiten, die hier aufgelistet sind:

(0111101111) (1011101111) (1011110111)
(1101101111) (1101110111) (1101111011)
(1110101111) (1110110111) (1110111011)
(1110111101) (1111001111) (1111010111)
(1111011011) (1111011101) (1111011110)

Nun kann man das Ganze wiederholen mit $k = 7$ usw. Es ist klar, dass […] Schwierigkeitsgrad und […] Zeitaufwand […] nicht angemessen sind."

Davies gibt keinen Algorithmus an, wie er die 15 Dualzahlen bei $k = 8$ Einsen gefunden hat. Davies (2009, S. 4) begründet das Vorgehen so:

„[…] (3) lässt mehrere Interpretationen zu. Ich analysiere drei. Da wir hierfür einen Wert für die ‚wahre' Wahrscheinlichkeit brauchen, werde ich trotz der oben angegebenen Kritik den Wert 0,904 einsetzen. […] Presse […] behauptet, dass die Anzahl der Versuche fehlt, um die Aufgabe lösen zu können. Da aber die Anzahl 10 in (1) und (2) angegeben wird, liegt es nahe, auch 10 in (3) anzunehmen."

Wie viele Dualzahlen (der Länge 10) gibt es, in denen keine Sequenz von 1en länger als 4 ist?

④ steht für eine Sequenz von (0) 1111 (0)
❹ analog für 0en
✹ notiert einen Block mit Nullen dazwischen

$k = 8$ Einsen: Anzahl erlaubter Sequenzen (also höchstens 4 Einsen in Serie) = 15, denn: Die 8 Einsen kann man in höchstens 3 Blöcke zerlegen (Summanden von höchstens 4; die kann man permutieren), denn man braucht jeweils einen 0er-Block zum Trennen und hat aber nur 10 – 8 = 2 Nullen. Die Blöcke mit Nullen muss man dann dazwischen (oder davor bzw. danach) einfügen und permutieren. Als Anzahl der möglichen Sequenzen erhält man $\binom{10}{8} = 45$.

Blöcke		Möglichkeiten		
1er	0er	1en	0en	alle
④✹④	❷	1	1	1
	❶, ❶		2	2
④✹③✹①	❶, ❶	3!	1	6
④✹②✹②	❶, ❶	3	1	3
③✹③✹②	❶, ❶	3	1	3

$k = 7$ Einsen: Anzahl der *verbotenen* Sequenzen (mindestens ein Block von 1en länger als 4) = 40; Anzahl der möglichen = $\binom{10}{7} = 120$; Anzahl der erlaubten = 120 – 40 = 80; denn:

Blöcke		Möglichkeiten		
1en	0en	1en	0en	alle
⑦	❸	1	2	2
	❷, ❶		2	2
⑥✹①	❸		1	2
	❷, ❶	2	2×2	8
	❶, ❶, ❶		1	2
⑤✹②	❸		1	2
	❷, ❶	2	2×2	8
	❶, ❶, ❶		1	2
⑤✹①✹①	❷, ❶	3	2×1	6
	❶, ❶, ❶		1×2	6

Die Fälle $k = 6$ und 5 sind noch überschaubarer; für $k \le 4$ sind alle Sequenzen erlaubt. Es wird nicht gesagt, dass die vorgestellte Zählung von den Maturanten eingefordert werden soll, aber sie entspricht durchaus dem Prinzip *gut* zu zählen. Wenn man kritisiert, dass eine Aufgabe zu schwierig ist, soll man sie nicht ganz unsystematisch lösen, damit man das noch unterstreicht. Als bedingte Wahrscheinlichkeiten erhält man im Fall k: $P(A \mid k)$ = Anzahl erlaubter / Anzahl möglicher Sequenzen (k); für $k = 8$ hat man: $P(A \mid k = 8) = 15/45 = ⅓$. Am Schluss muss man noch die Fälle $k = 0$ bis 10 nach der Binomialverteilung $B(10, 0{,}904)$ gewichten. Als Lösung ergibt sich $P(A) = 0{,}107$.

2. „Wir nehmen nun an, […] $n = 5$ ist. […] für diese Interpretation spricht […], dass die Lösung mit der Lösung des Ministeriums übereinstimmt. […] [Dieses schreibt] ‚Man betrachtet das Gegenereignis, dass er fünf Treffer hintereinander schafft […]'. […] Gegenereignis von "höchstens vier" ist nicht "fünf", sondern "mindestens fünf". Es ist nur dann fünf, wenn er genau fünfmal wirft. [Steht] nirgends in der Aufgabe und es gibt keinen Grund, dies anzunehmen. Wenn er nur fünfmal wirft, dann ist die Lösung des Ministeriums

$$1 - 0,904^5 = 1 - 0,6037 = 0,3963$$

sowie die Begründung korrekt. In allen anderen Fällen ist die Begründung falsch." (S. 4)

3. „[…] stellen wir uns vor, […] beim Training folgendes Spiel zwischen zwei Spielern :

Einer fängt an ([…] durch […] Münzwurf) und macht Freiwürfe bis zum ersten Fehlwurf. Der zweite Spieler ist nun an der Reihe […] bis zum ersten Fehlwurf usw. Nowitzki […] fängt an.

[…] Ws., dass er bei seinem ersten Versuch höchstens viermal erfolgreich war.

Die möglichen Versuchsfolgen sind

0 10 110
1110 11110

mit Wahrscheinlichkeiten

0,096 $0{,}904 \times 096$ $0{,}904^2 \times 0{,}096$
$0{,}904^3 \times 0{,}096$ $0{,}904^4 \times 0{,}096$

[…], dass die gewünschte Wahrscheinlichkeit

$$1-0{,}904^5 = 0{,}3963$$

beträgt. Man stellt fest, dass auch diese Lösung mit der Lösung des Ministeriums übereinstimmt. Die Begründung ist aber eine ganz andere. Die Interpretation ist die einzige, bei der die Anzahl der Würfe nicht von vornherein festgelegt ist. Ein sehr guter Schüler […]." (S. 4)

Für die Annahme $n = 5$ gibt es allerdings keine Begründung, auch nicht, dass die Lösung daraus mit der offiziellen übereinstimmt; die offizielle Lösung bekommt man nämlich *ohne* die Festlegung der Zahl der Versuche.

Mag sein, dass in der ministeriellen Lösung ein wenig salopp formuliert wurde; ‚mindestens fünf' muss genauer heißen ‚mindestens fünf und was weiter passiert, ist uns egal'.

Ist A das Ereignis ‚höchstens vier Erfolge', so lautet Gegenereignis $\overline{A}$ ‚fünfmal treffen *und es ist egal, was in den weiteren Versuchen passiert*'.

Wieso diese gekünstelte Spielsituation?

Mit $\Omega = \{0, 1\}$ ergibt sich $\overline{A}$ zu

$$(\underbrace{X_1 = 1, X_2 = 1, ..., X_5 = 1}_{=:S}, X_6 \in \Omega, X_7 \in \Omega, ...)$$

Klar:

$$P(X_6 \in \Omega, X_7 \in \Omega, ... \mid S) = 1 \cdot 1 \cdot ... = 1$$

$$P(\overline{A}) = P(S) \cdot P(X_6 \in \Omega, X_7 \in \Omega, ... \mid S) = p^5 \cdot 1$$

und $P(A) = 1 - p^5$.

Statt bei X_1 kann man auch bei X_i anfangen und die nächsten fünf Würfe entscheiden darüber, ob A oder $\overline{A}$ eintritt. Man muss nur vermeiden, den Beginn der Beobachtungen vom Ergebnis der Beobachtung abhängig zu machen.

Eine Bernoulli-Kette bleibt genau dieselbe Bernoulli-Kette (ihre grundlegenden Eigenschaften bleiben unverändert), wenn man den Anfangszeitpunkt beliebig wählt, oder wenn man aus der Kette durch Zufall einige Ergebnisse eliminiert.

Ohne dieses gekünstelte Spiel von Davies kann man zu einem echten Spiel einfach hingehen und – einmal da – beobachten, was passiert.

Fundamentale Eigenschaften von Bernoulli-Ketten

Wenn etwas fundamental an Bernoulli-Ketten ist, dann dieselbe Wahrscheinlichkeit für Erfolg in allen Versuchen, die Unabhängigkeit zwischen den Versuchen *und* die Beliebigkeit des Anfangszeitpunkts der Beobachtungen:

Wenn man einfach hinkommt und beobachtet, ob Nowitzki es schafft, mehr als viermal hintereinander zu treffen, dann ist alles über die fünfte Beobachtung hinaus überflüssig: die Lösung stimmt daher mit einer Vorgabe von $n = 5$ überein.

Wenn man das Modell einer Bernoulli-Kette unterstellt (was sehr wohl angreifbar ist), ist die Frage *ohne* die Angabe der Anzahl der Beobachtungen einfach zu lösen. Mathematisch ist die Aufgabe im Originaltext „wohldefiniert", denn die Lösung ist unabhängig vom Anfang i_0 der Beobachtungen:

$$(\underbrace{X_1 \in \Omega, X_2 \in \Omega, ..., X_{i_o-1} \in \Omega,}_{T} \underbrace{X_{i_o} = 1, X_{i_o+1} = 1, ..., X_{i_o+4} = 1,}_{S} \underbrace{X_{i_o+5} \in \Omega, X_{i_o+6} \in \Omega, ...}_{R})$$

Klar (die Beobachtungen sind stochastisch unabhängig):

$$P(T \cap S \cap R) = P(T) \cdot P(S \mid T) \cdot P(R \mid S \cap T) = 1 \cdot p^5 \cdot 1 \text{ und } P(A \mid \text{Beginn bei } i_0) = 1 - p^5 .$$

Man darf nur den Anfang der Beobachtungen nicht an den Daten orientieren: Wir beginnen zu beobachten, wenn gerade ein Erfolg da war und den zählen wir dann auch schon mit. Oder gar: Wir beginnen nach drei Erfolgen, zählen die mit und warten nur mehr den 4. und 5. Versuch ab.

Eine mit dem Originaltext verwandte Fragestellung sucht nach der Wahrscheinlichkeit des Ereignisses *B*, das aus allen Folgen von 0en und 1en besteht, in denen *nie* eine Sequenz von 1en auftaucht, die länger als vier ist. Da die Lösung dieser Aufgabe weitere Einsichten in Bernoulli-Ketten (und allgemeiner in Stichproben) bietet, sei auch sie hier skizziert:

Sei B_i das Ereignis, Nowitzki hat im *i*-ten Fünferabschnitt (seit Anbeginn; der erste Abschnitt umfasst Versuche 1-5, der zweite die Versuche 6-10 usw.) nicht mehr als vier Erfolge. Dann gilt $B \subseteq \bigcap_{i=1}^{n} B_i$, denn in einer Folge, in der nirgends eine Sequenz von mehr als vier 1en vorkommt, kommt auch *innerhalb* der ersten *n* Fünferabschnitte keine solche Sequenz vor. Damit erhält man wegen $P(B_i) = q$ mit $0 < q < 1$ ($q = 1 - p^5$!) und der Unabhängigkeit der B_i:

$$0 \le P(B) \le P(\bigcap_{i=1}^{n} B_i) = q^n \text{; daraus folgt } P(B) = 0 \text{, weil } n \text{ beliebig ist und } \lim_{n \to \infty} q^n = 0 \text{ gilt.}$$

Ganz allgemein gilt für Bernoulli-Ketten: Wenn man nur lange genug beobachtet, wird man auch die ungewöhnlichsten Ereignisse beobachten können; eine Serie (von 1en oder eine bestimmte Sequenz), sei sie auch noch so lang und daher selbst sehr unwahrscheinlich, wird irgendeinmal beobachtet werden können (denn $P(\overline{B}) = 1$).

Umformulierungen der Original-Aufgabe

An dieser Stelle sei an den Wortlaut der Aufgabenstellung von (3) erinnert: „Berechnen Sie die Wahrscheinlichkeit, dass er höchstens viermal nacheinander bei Freiversuchen erfolgreich ist."

Es mag verlockend sein, dem Text hinzuzufügen „… bei ***n*** Freiversuchen …". Das mag zusätzlich daher rühren, dass man so gewöhnt ist, die Zahl der Versuche vorgegeben zu bekommen, oder daher, dass man sich nicht so recht vorstellen kann, dass die Aufgabe ohne diesen Zusatz überhaupt sinnvoll ist. Stellt man so einen Bezug her, dann fehlt zunächst die Angabe von *n*, der Zahl der Beobachtungen.

Eine weitere Umformulierung des Texts liegt auch sehr nahe: „Berechnen Sie die Wahrscheinlichkeit, dass er ***immer*** höchstens viermal nacheinander bei Freiversuchen erfolgreich ist“ (d. h., nirgends ist eine Sequenz von 1en länger als 4). Wenngleich auch diese Formulierung die Aufgabenstellung völlig verändert, könnte man es einem Maturanten, der sie löst, als Teilleistung anrechnen. Als Lösung hat sich weiter oben 0 ergeben. Auch hier fehlt der Bezug auf die Zahl *n* der Beobachtungen; führt man einen ein, so erhält man eine unlösbare Aufgabe, solange man nicht die Zahl *n* explizit nennt.

Beide Umformulierungen führen zur selben Aufgabe, wenn man die Zahl der Beobachtungen vorgibt. Für $n = 10$ wurde sie gelöst; für größere Werte wird dies schwierig. Intuitiv kann man sagen: Je länger man den Spieler beobachtet, eine desto größere Chance erhält er, doch ‚mehr als viermal hintereinander' ($\overline{B} \mid n$) zu treffen – irgendwann in der Beobachtungszeit; es gilt: $P(B \mid n) \to 0$ für $n \to \infty$.

Eine zentral gestellte Aufgabe kommt einem juristischen Text gleich. Wenn man die Aufgabenstellung durch Texterweiterung *verändert*, kann es – wie hier – passieren, dass die neu entstehende Aufgabe unlösbar ist; in der Formulierung von (3) wurde kein Bezug zur Zahl der Beobachtungen hergestellt und konsequenterweise auch keinerlei Angabe über die Größe von *n* gemacht. Dem Prüfling Textveränderungen zuzugestehen, hieße aber, die Prüfung ad absurdum zu führen. Falls die Fragestellung im Original – und nicht irgendeine – zu einer unlösbaren Aufgabe führt, dann wäre allerdings Kritik angemessen. Die ursprüngliche Aufgabe *ist* aber lösbar, wie die Ausführungen weiter oben zeigen.

Es mag ja überraschen, dass die Aufgabe tatsächlich *ohne* den Bezug auf die Zahl der Beobachtungen Sinn hat und lösbar ist. Das darf jedoch nicht dem Aufgabenkonstrukteur zur Last gelegt werden. Man kann die „Überraschung“ vielleicht auch darauf zurückführen, dass die angesprochene Eigenschaft von Bernoulli-Ketten im Unterricht zu wenig (direkte) Beachtung findet. Bei der starken Ausrichtung auf das Rechnen *innerhalb* der Modelle mag die Erörterung zu kurz kommen, was diese Modelle eigentlich für die reale Situation vorgeben. Die angesprochene Eigenschaft von Bernoulli-Ketten trifft allgemeiner auch auf Stichproben zu und verdient sicherlich mehr Beachtung im Unterricht.

4. Ist Nowitzki auswärts schlechter als daheim?

Innerhalb des (kritisierten) Binomial-Modells wird gezeigt, wie man die Frage aus dem Kontext in ein statistisches Testproblem umsetzt. Dabei wird die von den Kritikern aufgeworfene Unterscheidung von wahren und geschätzten Parametern wieder aufgegriffen. Zum einen wird eine Saison als abgeschlossene Einheit aufgefasst, sodass man den Wert der Spielstärke genau kennt (und nicht schätzt). Zum anderen werden Verfahren vorgestellt, welche die unterschiedliche Behandlung der Spielstärke (einmal wird sie geschätzt, ein anderes Mal wird ihr Wert als Testgröße verwendet) vermeiden.

Das Nowitzki-Problem ist mit einem kleinen Dreh Stellvertreter für den Erkenntnisprozess in der empirischen Forschung. Argumente einer Homogenisierung werden thematisiert, welche eine Verletzung der Annahmen bei *statistischer* Betrachtungsweise als weniger gravierend erscheinen lassen. Zudem werden Confounder als ein wichtiges Instrument der praktischen Umsetzung von statistischen Tests dargestellt, welche die Homogenisierung der zu vergleichenden Gruppen besser umsetzen lassen als die Prüfung der Voraussetzungen einer Bernoulli-Kette (allgemeiner einer zufälligen Stichprobe).

Die zentral gestellte Aufgabe umfasst zwei Teile, die erste Frage war, ob Nowitzki auswärts schlechter ist als in allen Spielen; die zweite, ob er auswärts schlechter ist als daheim. Die erste ist einfach schlecht gestellt. Um mit Bildern zu sprechen: Man wird Unterschiede zwischen Gesunden und Kranken nicht dadurch beschreiben, indem man die Kranken mit *allen* vergleicht. Das sagt der Hausver-

stand. Also wird man wird die unterschiedliche Spielstärke auswärts und daheim – sofern sie besteht – nicht durch Unterschiede in den Treffern in Auswärtsspielen mit den Treffern in allen Spielen herausarbeiten. Da jedoch die Argumente von Davies sich auf diese schlecht gestellte Fragestellung von Auswärts gegen alle Spiele beziehen, behandeln wir diese Fragestellung wider besseres Wissen.

Trefferzahl statt Trefferquote – Ein Tippfehler

Die Kritik von Davies bringt eine merkwürdige nicht-statistische Betrachtungsweise eines statistischen Tests ein. Dem Schluss, dass die fehlerhafte Formulierung von *Trefferzahl* anstatt *Trefferquote* in der Aufgabenstellung schuld für die Sackgasse sei, in welche ihre Betrachtungsweise führt, muss entschieden begegnet werden. Die Aufgabe ist sowohl mit Trefferzahl als auch mit Quoten lösbar.

Wir müssen zwischen Heim- und Auswärtsspielen unterscheiden, daneben gibt es noch die Kategorie ‚alle Spiele'. Fiktiv könnten wir drei verschiedene Bernoulli-Ketten unterscheiden; für ihre Parameter siehe Tabelle 4. Man kann Untergruppen als *abgeschlossene* Einheiten betrachten; etwa können alle Spiele als Gesamtheit für sich angesehen werden und *nicht* als Realisierung von $n = 551$ Versuchen mit einer unbekannten Erfolgswahrscheinlichkeit p_g. Als Gesamtheit betrachtet ergibt sich ein Erfolgsscore von 498/551 = 0,904. Es erhebt sich natürlich die Frage, ob dieser Zahl eine Deutung als Wahrscheinlichkeit zugesprochen werden kann oder soll – siehe die Kritik weiter oben.

Tabelle 4: Bezeichnungen, Erwartungswerte und Parameter (Erfolgswahrscheinlichkeiten).

Spiele	Erfolge als Zufallsvariable	Zahl der Versuche	Beobachtete Erfolge	Erfolgswahrscheinlichkeit	Erwartungswert
Heim	T_H	n_H	k_H	p_H	$n_H \cdot p_H$
Auswärts	T_A	n_A	k_A	p_A	$n_A \cdot p_A$
Alle	$T = T_H + T_A$	$n = n_H + n_A$	$k_H + k_A$	p_g	$n \cdot p_g$

(K)Eine statistische Betrachtungsweise

„Nun, die wahren Werte p_H und p_A kennen wir nicht. Das Ministerium ersetzt sie einfach durch die Schätzwerte $\hat{p}_H$ und $\hat{p}_A$ […]

‚Es sei X B263; 0,904–verteilt'

[… woraus die] Verwechslung eines Schätzwertes mit dem wahren Parameterwert sichtbar ist. [Ist] der "Erwartungswert" für Heim- und Auswärtsspiele nun nichts anders als

$T_H + T_A$.

Die Trefferanzahl bei Auswärtsspielen ist T_A und wir müssen nun testen,

ob T_A signifikant kleiner ist als $T_H + T_A$.

Dies ist nun ja immer der Fall, […] Setzt man die Daten ein, müssen wir testen, ob

$231 < 267 + 231$, was ja stimmt.

[…] vermutlich ein Tippfehler […] nicht die "Trefferanzahl", sondern die Trefferquote." (S. 5)

Statistischer Vergleich von Daten

Es hat keinen Sinn zu fragen, ob p_A signifikant kleiner ist als p_g. Man kann nicht einfach zwei *Zahlen* ‚statistisch' miteinander vergleichen.

Wird die Saison als Einheit gesehen, kennt man die Spielstärke, d. h. $p_g = 0{,}904$; man kann fragen: Ist es ‚vernünftig', die Auswärtsspiele als wiederholtes Bernoulli-Experiment mit eben dieser Wahrscheinlichkeit 0,904 zu modellieren?

Die Daten $T_A = 231$ in $n_A = 263$ Auswärtsspielen werden mit $B(n_A = 263, \pi = 0{,}904)$ als Referenzverteilung verglichen; diese Nullhypothese beschreibt die Verteilung der Treffer, wenn für die Auswärtsspiele eine Spielstärke von 0,904 zutrifft. Der Erwartungswert für die Daten ist

$E(T_A \mid 0{,}904) = 263 \times 0{,}904 = 237{,}7$

und nicht $267 + 231$.

$T_A = 231$ kann aber nicht mit dem Erwartungswert 237,7, sondern muss mit dem 5%-Quantil der Referenzverteilung verglichen werden.

Die Sprechweise der Aufgabenstellung „ob die Trefferzahl bei Auswärtsspielen signifikant unter dem *Erwartungswert* für Heim- und Auswärtsspiele liegt“ ist irreführend. Gemeint ist: Sind die Daten der Auswärtsspiele damit vereinbar, dass Nowitzki auswärts eine Spielstärke wie in allen Spielen aufweist, oder sprechen die Daten (signifikant) dagegen? Die Daten sprechen dagegen, wenn sie aus dem ‚Rahmen' der normalen Zufallsschwankungen herausfallen; diese Schwankungen beziehen sich auf die Voraussetzung, die Spielstärke p_g aller Spiele trifft auch für Auswärtsspiele zu. Salopp formuliert, wenn die Daten mit dem, was normalerweise unter p_g passiert, nicht mehr ‚zusammenpassen'. Jede Verkürzung jedoch birgt die Gefahr von Missverständnissen. Was normalerweise zu erwarten ist, ist ein *Spielraum von Werte*n, aber nicht der Erwartungswert – auch wenn das sprachlich ganz knapp beieinander liegt. Die Notation wurde gegenüber dem Original ein wenig geglättet; so wurde u. a. $T_{n_H}^H$ durch T_H ersetzt; der Bezug zu der Anzahl der Daten ergibt sich ohnehin aus dem Zusammenhang.

Umsetzung der Fragestellung in ein statistisches Modell

Das merkwürdige Argument von Davies, das in die Sackgasse für die Trefferzahlen führt, ist Anlass dafür, die Denkweise eines statistischen Tests pointiert nocheinmal wieder zu geben. Es geht um die Bewertung von Daten über Auswärtsspiele in einer geeigneten Referenzverteilung. Verglichen werden die Daten natürlich nicht mit dem Erwartungswert der Referenzverteilung sondern mit Quantilen, die den Kern der Verteilung von extremen Außenbereichen abgrenzen. Fallen die Daten aus dem ‚normalen' Spiel des Zufalls heraus, überschreiten sie also diese Schwellenwerte oder Ablehnzahlen, so deutet dies darauf hin, dass sie mit der Referenzverteilung nicht vereinbar sind.

Die Fragestellung ‚Ist Nowitzki in Auswärtsspielen bei seinen Freiwürfen schlechter als in *allen* Spielen der Saison?' setzt man also wie folgt in ein statistisches Testproblem um:

Wahl des statistischen Modells

$$T_A \sim B(n_A, \pi),$$

d. h., die Trefferzahl in Auswärtsspielen ist binomial verteilt mit unbekanntem Parameter π. Wir wollen die Voraussetzungen der Bernoulli-Kette jetzt nicht noch einmal hinterfragen.

Wahl der Hypothesen gemäß der Sachfrage

Als Nullhypothese wählen wir $\pi = p_g$, wobei p_g die Spielstärke in allen Spielen darstellt. Als Alternativhypothese wählen wir die einseitige Hypothese $\pi < p_g$.

Bewertung der Daten

Wir vergleichen die Daten über die Auswärtsspiele mit dieser Hypothese. Liegen sie in der Mitte dieser Verteilung oder an den Rändern? Je weiter ‚draußen' sie liegen, desto größer der ‚Einwand', dass sie dazu gehören. Liegen die Daten im extremen Bereich der damit festgelegten Verteilung, dann lehnen wir die Nullhypothese ab und sagen: die festgestellten Abweichungen von der Nullhypothese sind signifikant (auf dem Niveau α) – sie können zwar auftreten, aber mit dem reinen Spiel des Zufalls nicht ausreichend erklärt werden.

Entscheidung

In einer Abwägung von ‚etwas Seltenes ist eingetreten' und ‚die Nullhypothese trifft doch nicht zu' (die Alternative trifft also zu), d. h., ‚Nowitzki ist auswärts schwächer als in allen Spielen' entscheidet man dann für die Alternative. Nicht, dass die Alternative dann *zutrifft*; man *entscheidet* sich für sie. Man könnte auch einen Fehler machen. Das Risiko wird mit α beziffert (hier wird nur der so genannte Fehler 1. Art erfasst, es gibt ja noch den Fehler 2. Art). Das Niveau α bestimmt die Ablehnzahl.

Über die Problematik ein- und zweiseitigen Testens gibt es viel zu sagen; außer in wohl begründeten Fällen, in denen man über die ‚Wirkung' sehr genau Bescheid weiß, sollte man eine einseitige Alternative vermeiden. Einseitige Alternativen führen leichter zur Ablehnung der Nullhypothese, was als statistischer ‚Nachweis' der Alternative gewertet wird. Die unbegründete Wahl einer einseitigen Alternative führt also leichter zum ‚Nachweis' dieser Alternative – diese methodologische Falle sollte man vermeiden. Im Sport ist es jedoch allgemein bekannt, dass es einen Heimvorteil gibt.

Schätzen der Spielstärke in allen Spielen aus den Daten

Sofern man die Spielstärke p_g in allen Spielen kennt, ist die Berechnung der Ablehnzahl und die Testentscheidung eine reine Frage der Routine. Im Allgemeinen aber muss man zur Festlegung des Parameters in der Nullhypothese doch noch Wissen aus dem Kontext heranziehen: Langjährige Erfahrung, Normwerte, deren Einhaltung gefordert ist, oder eine Schätzung aus vergangenen Daten.

i) Die Nullhypothese als Referenzverteilung für die Daten zu Auswärtsspielen fußt auf einer Schätzung, welche Fehler beinhaltet. Um der daraus resultierenden Ungenauigkeit zu begegnen, müsste man – wie im Wahrscheinlichkeitsteil – eigentlich ein Konfidenzintervall für p_g bestimmen; bei 95% Sicherheit erhält man das Intervall: [0,879, 0,928]. Der ungünstigste Fall für alle Spiele im Sinne der Fragestellung, ob Nowitzki in Auswärtsspielen schwächer ist als in allen, ist demnach $p_g = 0{,}879$.

Mit diesem Wert müsste man die Referenzverteilung aus der Nullhypothese aufbauen und nicht mit $\hat{p}_g = 0{,}904$. Das 5%-Quantil dieser Verteilung ist 222,5. Der Wert von $T_A = 231$ liegt nicht darunter, was zu keiner Ablehnung der Nullhypothese führt.

Die Interpretation so eines Verfahrens ist aber deutlich schwieriger. Es ist ja für die meisten Schüler schon schwierig genug, die Bedeutung eines normalen Signifikanztests in (eigene, richtige) Worte zu fassen. Aus ähnlichen Gründen hat sich das Verfahren auch in der Praxis nicht durchgesetzt; man schätzt zwar Parameter, um Modelle zu fixieren, rechnet aber immer mit *einem* Modell, ohne den Schätzfehler zu berücksichtigen. Bei den alternativen Verfahren (siehe den folgenden Unterabschnitt) wird das Problem umgangen.

Die Lösung basiert auf fiktiven, unterstellten Werten für die Spielstärke in *allen* Spielen, die allesamt nicht den wahren Wert dafür darstellen. Mit diesem fiktiven Wert modelliert man die Freiwürfe auswärts als Bernoulli-Kette und berechnet die Konsequenzen für die Wahrscheinlichkeitsverteilung der Treffer. Dann reiht man die vorhandene Beobachtung in diese Verteilung ein und beurteilt, ob sie dazu passt oder nicht. Die fiktiven Werte entsprechen einem Wissen über alle Spiele. Dieses Wissen ist weiter oben als faktisch unterstellt; hier wurde es aus Daten geschätzt, wobei die auftretende Ungenauigkeit dadurch ‚aufgefangen' wurde, dass man ein Konfidenzintervall berechnet (und daraus allenfalls den schlechtesten Fall herangezogen) hat.

ii) Wenn man den Parameter aus denselben Daten schätzt, an denen man dann den Test durchführt, macht man methodisch einen Fehler. Regel: Man formuliere nie die Nullhypothese aus denselben Daten, an denen man sie testet. Im extremen Fall würden dieselben Daten zur Schätzung der Spielstärke herangezogen, an denen die zugehörige Nullhypothese über die Spielstärke getestet wird; damit könnte man die Nullhypothese nie ablehnen. In der schlecht gestellten Frage ‚Auswärts schlechter als in allen Spielen?' kommt es teilweise zu so einer Vermischung.

Wenn man die Frage anders stellt als ‚Auswärts schlechter als daheim?', so wird die Nullhypothese als $\pi = p_H$ formuliert. Dann gibt es keine Überschneidung von Daten mehr, die zur Schätzung des Parameters aus der Nullhypothese führen, und den Daten, an denen getestet wird. Übrig bleibt allerdings

der Fehler, der bei einer solchen Schätzung passieren kann. Um das zu berücksichtigen, kann man wieder ein Konfidenzintervall heranziehen.

Im Sinne der einseitigen Fragestellung ‚Auswärts schlechter als daheim?' ist der ungünstigste Fall, der für Heimspiele in Frage kommt, der untere Rand dieses Konfidenzintervalls. Eine solche Betrachtung wird zumeist nicht durchgeführt, obwohl sie – siehe die große Marge in den Ergebnissen im Wahrscheinlichkeitsteil der Aufgabe – sinnvoll wäre.

„Der Erwartungswert der Trefferquote für Heim- und Auswärtsspiele beträgt

$$p_g = \frac{n_H}{n_g} \cdot p_H + \frac{n_A}{n_g} \cdot p_A .$$

[…] anhand der Daten testen, ob

$$p_A < p_g = \frac{n_H}{n_g} \cdot p_H + \frac{n_A}{n_g} \cdot p_A .$$

Da aber die wahren p_H und p_A nicht bekannt sind, hat das Ministerium sie in […] p_g einfach durch die Schätzwerte $\hat{p}_H$ und $\hat{p}_A$ ersetzt und tat so, als ob sie die wahren Werte sind […]

Gleichzeitig wird aber in der Lösung des Ministeriums $\hat{p}_A$ als Schätzwert für p_A behandelt. […], $\hat{p}_A$ wird manchmal als der wahre Wert betrachtet und manchmal als ein Schätzer hierfür und noch dazu im selben Satz. Die Begriffsverwirrung ist komplett." (S. 5)

Es stört tatsächlich die unterschiedliche Behandlung von p_A und p_H, weil ja p_g sich als Konvexkombination der beiden ergibt. Da werden Teile von p_g einmal geschätzt, einmal nicht geschätzt. Der Vergleich von Auswärtsspielen gegen alle Spiele entspricht aber einer falsch gestellten Frage. Eigentlich müssten wir ja Auswärts- gegen Heimspiele vergleichen.

Vergleicht man Auswärtsspiele nur gegen Heimspiele, so hat man in der ‚Logik' von Davies

$$p_A < p_H$$

zu testen und müsste wieder einmal zu Schätzungen greifen für p_H, das andere Mal aber die Daten in Auswärtsspielen (die ja analog p_A schätzen ließen) nur als Wert einer Teststatistik begreifen, die man mit einem kritischen Wert aus der Nullhypothese zu vergleichen hat.

Der Fragestellung, ob Nowitzki auswärts schwächer ist als daheim, entspricht tatsächlich die Formulierung folgender Nullhypothese für die Auswärtsscores: $T_A \sim B(n_A, p_H)$ und ein anschließender Vergleich der Daten T_A für die Auswärtsspiele: Ist der Wert 231 zu klein für diese Verteilung oder nicht? Liegt 231 im unteren Ablehnungsbereich (bei einseitiger Alternativhypothese $p < p_H$)?

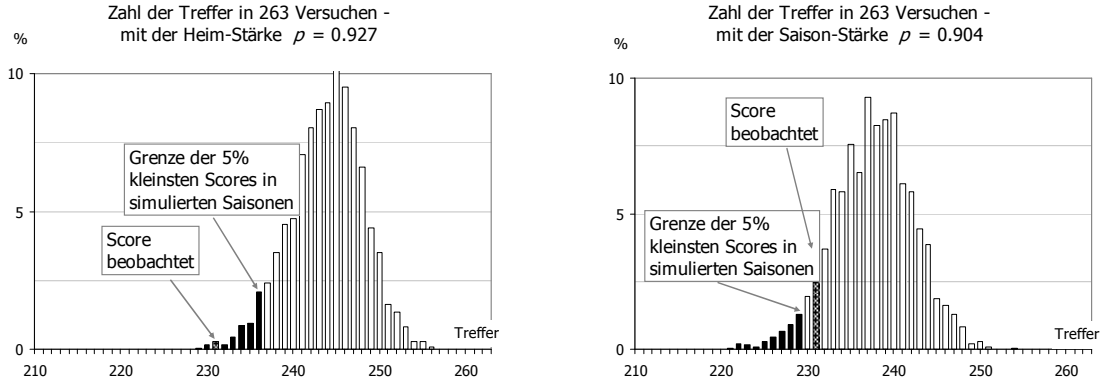

Abb. 2: Ergebnisse von 2000 fiktiven Saisonen mit n_A = 263 Versuchen –
Links mit einer unterstellten Stärke von p = 0,927 (entsprechend den Heimspielen) – Rechts mit p = 0,904 (alle Spiele).

Auch hier kann man das Wissen über die Heimstärke faktisch aus den Daten der gesamten Saison ablesen. Oder man kann die Heimstärke durch die Daten $\hat{p}_H$ schätzen. Jetzt gibt es keine Überlappung mit Daten, die zur Festlegung der Nullhypothese führen, und jenen Daten für die Auswärtsspiele,

mit denen wir den Test durchführen. Auch hier müssten wir den Ungenauigkeiten der Schätzung $\hat{p}_H$ durch ein Konfidenzintervall Rechnung tragen und die Nullhypothese als ungünstigsten Fall formulieren. Man kann aber auch die Lösung mit der – nach der Saison – faktisch bekannten Wahrscheinlichkeit für die Heimspiele vertreten und auf diese Verkomplizierung verzichten.

Alternativen zur Beantwortung der Kernfrage

Zur Beantwortung der Frage ‚Ist Nowitzki auswärts schlechter als daheim?' gibt es mehrere Verfahren. Diese werden hier vorgestellt. Alle umgehen die unterschiedliche Behandlung der Parameter. Diese ist zum Teil dadurch verursacht, dass bisher der Apparat an statistischen Tests auf Ein-Stichproben-Methoden eingeschränkt war. Diese Diskussion um die Schätzung oder Nicht-Schätzung von Parametern entstand also primär daraus, dass man von Seiten der Aufgabensteller eine interessante Frage aus einem motivierenden Kontext zur Prüfung stellen wollte. Allerdings fehlen im schulischen Vorrat an statistischen Tests Verfahren zu Zwei-Stichproben-Problemen, obwohl einige durchaus mit den üblichen Schulkenntnissen erschließbar wären.

Die ‚Haarspaltereien' mit der unterschiedlichen Behandlung der Parameter p_A und p_H, in die man kommt, wenn man den Ausweg mit der nach der Saison bekannten Wahrscheinlichkeit nicht ‚anerkennt', kann man natürlich vermeiden. Einige Möglichkeiten führen über das Niveau der Matura hinaus (siehe Borovcnik, 2009, oder Borovcnik und Kapadia, 2011): Der χ^2-Test auf Unabhängigkeit der Merkmale ‚Erfolg' in einem Freiwurf und ‚Ort des Freiwurfs' (daheim bzw. auswärts) kommt ohne solche Schätzungen aus. Der Test der Nullhypothese $p_A - p_H = 0$ gegen die Alternative $p_A - p_H < 0$ mittels einer geeigneten Standardisierung der Testgröße $\hat{p}_A - \hat{p}_H$ behandelt das Problem als Zwei-Stichproben-Problem (zwei Bernoulli-Ketten mit ihren jeweiligen Erfolgsparametern) und schätzt beide Parameter – behandelt sie also gleich; die technischen Details sind aber in der üblichen mathematischen Behandlung der Begriffe und Methoden für den Schulunterricht (und nicht nur dafür) zu komplex. Allerdings stellt das Verfahren einen wichtigen Meilenstein in der empirischen Forschung dar, sodass es sich lohnen würde, es unter Modellierungsgesichtspunkten doch in den Unterricht aufzunehmen. Eine gute mathematische Erklärung der Verfahren ist in Mosler und Schmid (2006) zu finden; für eine Darstellung aus Anwendersicht siehe Lorenz (1996).

Ein feiner Ausweg, der in gewisser Hinsicht der Modellierung mit zwei Bernoulli-Ketten überlegen ist, hängt mit der hypergeometrischen Verteilung zusammen. Die Daten werden als endliche Gesamtheit für sich betrachtet und nicht als Beobachtung von (unendlichen) Bernoulli-Ketten. Nur zufällig gezogen muss auch beim Fisher'schen exakten Test werden.

Tabelle 5: Anzahl der Treffer und Versuche.

Spiele	Treffer	Nieten	Versuche
Heim	267	21	288
Auswärts	231	32	263
Alle	498	53	551

Alle 551 Versuche sind als Kugeln in einer Urne repräsentiert; 498 Kugeln sind ‚weiß' (Erfolg) und 53 ‚schwarz' (Niete). Man hat nun für die Auswärtsspiele 263 Ziehungen aus dieser Urne und erhält 231 weiße. Wenn zu wenig weiße gezogen werden, ist das ein klarer Einwand gegen die Nullhypothese, dass für die Auswärtsspiele aus dieser Urne gezogen wird.

Wie wahrscheinlich sind 231 oder gar weniger weiße Kugeln?

Die Antwort erhält man aus der hypergeometrischen Verteilung mit den genannten Parametern. Die Ablehnzahl ist dann das 5%-Quantil dieser Verteilung. Wenn 231 kleiner als diese Ablehnzahl ist, so wird die Nullhypothese, dass Nowitzki in Auswärtsspielen gleich gut ist wie in allen Spielen, abgelehnt. Alternativ kann man den so genannten *p*-Wert (*p* value, hat nichts mit dem *p* der Binomialverteilung zu tun) der Beobachtung 231 bestimmen; das beantwortet die oben gestellte Frage direkt. Es ergibt sich ein *p*-Wert von 3,6%; damit kann man – auf dem 5%-Signifikanzniveau (einseitig) die Hypothese der Gleichheit der Spielstärken ablehnen.

Es wurden hierbei weder die Spielstärken für Heim- noch Auswärtsspiele (auch nicht bei allen Spielen) geschätzt. Die Lösung kommt ohne solche Schätzwerte aus und beantwortet die Sachfrage, ob man davon ausgehen kann, dass Nowitzki auswärts schwächer ist als daheim.

Schätzwerte und feste Größen

Die Unterscheidung von Schätzwerten und wahren Werten ist – nach Davies – von grundlegender Bedeutung. Während die grundlegende Bedeutung nicht angezweifelt wird, wird die Hervorhebung von *wahren Werten* als hinderlich für den Aufbau eines geeigneten Verständnisses für die beurteilende Statistik erachtet. Man hat fiktive Werte zu unterstellen und testen. Wie man diese fiktiven Werte bekommt, entspricht oft den Fragestellungen und dem Wissen aus dem Kontext, das man in fiktive Werte über einen Parameter umsetzen kann oder nicht.

Sind die beobachteten Trefferzahlen (oder Trefferquoten, das ist gleichwertig) im Ablehnungsbereich der Nullhypothese oder nicht? Dabei entspricht die Formulierung der Nullhypothese der Festlegung einer Referenzverteilung, die der Sachfrage entspricht. Lautet die Frage im Kontext ‚Ist Nowitzki auswärts schwächer als daheim?', so ist der Erfolgsparameter für die Auswärtsspiele mit p_H gleich zu setzen. Dieser Wert kann nach Ende der Saison als bekannt angesehen und daher aus den Daten abgelesen (nicht geschätzt) werden. Wenn man diesen Wert aber als unbekannt auffassen will und ihn aus den Daten schätzt, so müsste man eigentlich die Ungenauigkeit der Schätzung über ein Konfidenzintervall kontrollieren. Egal, ob man das berücksichtigt oder (‚fälschlicherweise') nicht, man erhält daraus fiktive Werte für die Nullhypothese zum Vergleich.

Die beobachtete Erfolgsrate in Auswärtsspielen ist eine empirische Größe, deren beobachteter Wert mit der Bezugsverteilung aus der Nullhypothese verglichen wird. Der Status von Schätzwert und feste Größe ist veränderbar – je nach Hypothese und je nach Ergänzung von fehlendem Wissen über die Parameter. Zwangsläufig kommt es zu einer ‚Vermischung' dieser Eigenschaften.

Voraussetzungen einer Bernoulli-Kette – probabilistische und statistische Anwendungen

In diesem Abschnitt wird erörtert, dass der Grad an Verletzung von Annahmen aus einer globalen statistischen Betrachtungsweise – merkwürdigerweise – weniger gravierend ist als aus einer probabilistischen. Beurteilt man eine kleine Sequenz von Freiwürfen, so sind die immer als besonders zu betrachten: die Form des Spielers, der Spielstand etc. weichen von der globalen Trefferwahrscheinlichkeit ab. Beim Vergleich von ganzen Blöcken wie Auswärts- gegen Heimspiele gibt es auch Abweichungen, aber – so argumentiert man tatsächlich – die Verletzung der Voraussetzungen wirkt sich auf die zu vergleichenden Blöcke ähnlich aus. Dann beeinträchtigen sie den Vergleich auch weniger.

In der Beurteilung, ob Nowitzki in Auswärtsspielen schlechter ist als in Heimspielen, ist das Szenario einer Bernoulli-Kette aussagekräftiger als bei der Berechnung von einzelnen Wahrscheinlichkeiten für kurze Abschnitte von Versuchen. Tatsächlich werden ganze Blöcke von Spielen gegeneinander vergli-

chen und es geht nicht um die Frage nach Wahrscheinlichkeiten für bestimmte Anzahlen in einzelnen Unterabschnitten, sondern um die Frage, ob es *Unterschiede* zwischen den Blöcken *im Gesamten* gibt.

Dabei wird die angesprochene Homogenisierung, das ist ein Ausgleich von irgendwelchen Abhängigkeiten zwischen einzelnen Versuchen, oder von unterschiedlichen Trefferquoten über einzelne Phasen einer Saison hinweg, eher wirksam.

Dieser Ausgleich von Effekten über einzelne Elemente hinweg auf die Gesamtheit ist ein grundlegender Bestandteil *statistischer* Betrachtungsweise.

Für eine statistische Betrachtung des Problems aus dem Kontext kann die Unterstellung eines Szenarios einer Bernoulli-Kette (mit derselben Erfolgswahrscheinlichkeit) durchaus zu relevanten Ergebnissen führen. Für eine alternative Betrachtung von Modellen als Szenarien, welche – im Gegensatz zu Modellen nicht unbedingt auf eine möglichst perfekte Anpassung an die Wirklichkeit abzielt, sei auf Borovcnik (2009 oder 2011) verwiesen. Wir wissen schon, eigentlich sind die Voraussetzungen der gleichen Erfolgswahrscheinlichkeit verletzt (Schwankungen der Form, Streits im Team etc.), ebenso fehlt die Unabhängigkeit der Versuche (Pechserien, ein guter ‚Lauf' etc.).

Dennoch wenden wir ein heuristisches Argument der Homogenisierung an, weil sich die Verletzungen der Annahmen vielleicht doch irgendwie ausbalancieren, jedenfalls aber – so das Argument – sich auswärts ähnlich ausprägen wie in den Spielen daheim. Das lässt uns die Szenarien von Bernoulli-Ketten anwenden und testen, ob die Erfolgswahrscheinlichkeit auswärts mit der in Heimspielen verträglich ist. Die Situation scheint viel besser zu sein als in der probabilistischen Anwendung, wo man kaum Situationen finden konnte, wo die Annahmen einer Bernoulli-Kette passen sollten; am ehesten noch in der künstlichen Situation, in der man sich aus den Videos von allen Freiwürfen zufällig 10 auswählt und nachsieht, ob Nowitzki höchstens acht Mal trifft.

Man kann versuchen, das Vorliegen der Voraussetzungen statistisch zu prüfen. Die gleiche Trefferwahrscheinlichkeit etwa. Wir haben hier gesehen, dass das selbst für so auf der Hand liegende Unterabschnitte von Freiwürfen in Heim- und Auswärtsspielen schwierig ist. Und das Ergebnis lautet: es gibt Unterschiede. Welche anderen Abschnitte soll man noch überprüfen? Was die Unabhängigkeit betrifft, so gibt es keine Möglichkeit, das Vorliegen einer Nullhypothese statistisch zu bestätigen. Die Rationalität von statistischen Tests hängt entschieden davon ab, dass wir auch einen Fehler 2. Art und die „Macht" (die Gegenwahrscheinlichkeit davon) berechnen können; die Macht ist eine Funktion von der Alternativhypothese und gibt an, wie wahrscheinlich es ist, die Nullhypothese abzulehnen, wenn eine spezielle Alternative statt ihrer zutrifft.

Geht es um ein parametrisches Problem (wird ein Parameter einer Modellfamilie getestet), kann man die Alternativen nach der Größe des Parameters anordnen und beurteilen, wie weit ein spezieller Wert der Alternative von der Nullhypothese entfernt ist. Etwa im Testproblem $\pi = \pi_0$ gegen $\pi \neq \pi_0$ kann man ermessen, wie ‚weit entfernt' ein spezieller Wert π von der Nullhypothese π_0 ist: die Stärke 0,85 liegt im Sportkontext weit unter 0,904. Eine Macht von z. B. 0,70 (für $\pi = 0{,}85$) besagt: der angewendete Test hat eine Wahrscheinlichkeit von 0,70, den Wert 0,904 aus der Nullhypothese abzulehnen, falls tatsächlich 0,85 zutrifft.

Solche Überlegungen fehlen bei einem Test auf Unabhängigkeit. Was macht man mit dem Ergebnis, die Nullhypothese der Unabhängigkeit konnte *nicht* abgelehnt werden? So wünschenswert dann eine Aussage ‚die Nullhypothese trifft zu' wäre, sie ist unbegründet. Selbst ein ‚sie trifft wahrscheinlich zu' entbehrt jeder Begründung. Das ist ein bekanntes Manko statistischen Testens, weswegen man die

‚erwünschte' (nachzuweisende Hypothese) immer als Alternative und das Gegenteil davon als Nullhypothese wählt.

Die Prüfung dagegen, ob bestimmte Abhängigkeiten zutreffen, ist wiederum der Willkür ausgesetzt, dass man irgendwelche Abhängigkeiten formulieren kann. Abhängigkeiten etwa als Übergangswahrscheinlichkeiten von ‚nach 3 Erfolgen ist ein 4. Erfolg um … weniger wahrscheinlich', sind im Sport wenig interessant. Das würde Abhängigkeiten wieder in einem stabilen Muster sehen. Auch nichtparametrische Ansätze zur Prüfung der Unabhängigkeit wie Run-Tests basieren auf konstanten Erfolgswahrscheinlichkeiten und bieten nur *heuristische* Hilfen zur Beurteilung der Verletzung der Annahmen. Erfolgversprechender ist es, nach Confoundern zu suchen statt Abweichungen von der Unabhängigkeit zu prüfen. Confounder sind etwa Ort des Freiwurfs (auswärts oder daheim), Formschwankungen (wie immer man die misst, jedenfalls umfassender als durch Zählen von ein paar Fehlwürfen), Verletzungen etc.

Empirische Forschung – Verallgemeinerung von Ergebnissen aus eingeschränkten Daten

In diesem Abschnitt wird eine Verbindung zur empirischen Forschung hergestellt. Diese soll aufzeigen, dass die angesprochenen Denkweisen generell wichtig sind. Bei der Verallgemeinerung von Wissen aus Daten kommt es immer wieder zu einer Vermischung von probabilistischen und heuristischen Argumenten, wieso das Modell passen soll. Das einfachste Problem ist der Vergleich von zwei Gruppen. Sind die beiden Gruppen hinsichtlich eines Zielmerkmals verschieden oder entbehrt es einer (empirischen) Grundlage, von solchen Unterschieden zu sprechen? Der Vergleich der Auswärts- und Heimspiele von Nowitzki bildet einen guten Einstieg. Fragen und Probleme sowie Methoden und heuristische Argumente ähneln sehr.

Man kann nicht einfach Daten per se interpretieren. Daten bekommen erst einen Sinn, wenn man sie im Lichte des Kontexts *und* von Modellen interpretiert. Wissen aus dem Kontext beeinflusst sowohl die Suche nach potentiellen Confoundern als auch die Interpretation und die Bewertung der praktischen Relevanz von Schlüssen, die man aus dem Modell zieht.

Ein Argument der Homogenisierung wurde herangezogen, um die Anwendung eines probabilistischen Modells zu rechtfertigen. Für Bernoulli-Ketten sollten sich die unterschiedlichen Erfolgswahrscheinlichkeiten in größeren Blöcken ausebnen, ebenso besteht die ‚Hoffnung', dass die Verletzungen der Unabhängigkeit aus der globaleren statistischen Sicht weniger relevant werden. Man ist aber vielleicht geneigt, diese Homogenisierung im Kontext des Sports als unrealistisch abzulehnen. Dabei ist die Situation ziemlich ähnlich der Standardsituation in der empirischen Forschung. Ein kleiner Dreh und wir erhalten, mit denselben Daten wie in der Nowitzki-Aufgabe, eine neue Situation – siehe Tabelle 6.

Tabelle 6: Wahrscheinlichkeiten für den Erfolg einer Behandlung in Versuchs- und Kontrollgruppe.

Gruppe	Erfolg	Anzahl	Erfolgswahrscheinlichkeit
Versuch	267	$n_T = 288$	$p_T = 0{,}927$
Kontrolle	231	$n_C = 263$	$p_C = 0{,}878$
Alle	498	$N = 551$	$p_g = 0{,}904$

Jetzt geht es um ein Zwei-Stichproben-Problem: Wir bewerten den Erfolg einer medizinischen Behandlung auf einer 0,1-Skala (nicht mit einem stetigen Zielmerkmal). Die Personen der Kontrollgruppe haben nur eine Scheinbehandlung (Placebo) erhalten, die Personen der Versuchsgruppe wurden mit dem zu testenden Medikament behandelt. Ist die Behandlung wirksam? Genauer: Ist die Behandlung wirksamer als Placebo? Natürlich können wir die Methoden aus dem Nowitzki-Problem anwenden.

Wie können wir die statistische Betrachtungsweise rechtfertigen? Wir müssen den Erfolg durch eine Bernoulli-Kette modellieren und fragen, ob beide Gruppen durch dieselbe Erfolgswahrscheinlichkeit ausgezeichnet sind. Dieses Modell erfordert ein und dieselbe Erfolgswahrscheinlichkeit für alle Personen, wenigstens für alle Personen ein und derselben Gruppe. Natürlich kommt auch die Unabhängigkeit des Erfolgs zwischen verschiedenen Leuten als Voraussetzung hinzu.

Solch ein probabilistisches Modell wird üblicherweise mit dem Design der Studie begründet. Natürlich werden die Leute alles andere als zufällig aus einer größeren Population ausgewählt; sie werden nach ‚Bequemlichkeit akquiriert' – sie sind hauptsächlich Patienten der Ärzte, welche die Studie durchführen. Man ordnet sie jedoch durch Zufall einer der beiden Gruppen zu, d. h., ein Zufallsexperiment entscheidet, welche Behandlung sie erhalten (Medikament oder Placebo). Die Placebo-Behandlung sieht äußerlich ganz genau so aus wie die echte Behandlung. Sie hat aber keine zu erwartende Wirkung, außer der psychologischen Erwartungshaltung der damit behandelten Menschen.

Es wurde heftig diskutiert, wie stark Placebo-Effekte tatsächlich sind; aber allein der Umstand, dass sich durch die Erwartungshaltung einer Behandlung eine Wirkung einstellen kann, lässt dies als Confounder ansehen und man muss danach trachten, diesen Confounder auszuschalten. Weder der Patient noch der behandelnde Arzt (oder Personen, welche medizinische Messungen für den Erfolg der Behandlung durchführen, oder die Personen betreuen) dürfen wissen, wer welche Behandlung wirklich erhält – der ‚golden standard' in der empirischen Forschung ist das so genannte doppelblinde randomisierte Experiment mit Versuchs- und Placebo-Kontrollgruppe. Aus ethischen Gründen muss man natürlich davon oft abweichen, etwa kann man bei Schwerkranken kein Placebo verabreichen, sondern wird die Kontrollgruppe mit dem bisherigen medizinischen Standard behandeln.

Die zufällige Zuordnung der Personen zu Behandlung bzw. Placebo soll die beiden Gruppen so vergleichbar wie möglich machen – alle bekannten Kovariaten und unbekannten Confounder sollen sich gleichmäßig auf die Gruppen auswirken; der Unterschied, der noch verbleibt, ist die Behandlung. Das heißt, wenn noch Unterschiede bestehen, so können sie auf die Behandlung zurückgeführt werden. Das entspricht wieder einem „ceteris paribus"-Argument: Wenn alle anderen Einflussfaktoren bis auf die Behandlung gleich sind, dann sind beobachtete Unterschiede im Zielmerkmal auf die Behandlung zurückzuführen.

Es ist dennoch anzuraten, dem Zufall nachzuhelfen. Das geht auf grundsätzlich zwei Arten: für die erste kennt man schon potentielle Einflussmerkmale. Wenn man z. B. vermutet, die Behandlung wirkt sich geschlechtsspezifisch aus, so wird man sicher stellen, dass unter Frauen und unter Männern *getrennt* zufällig zu den Gruppen zugeordnet wird. Die zweite ist, man weiß um mögliche Einflussfaktoren, aber eine Schichtung ist schwer möglich. Man zeichnet die Daten über diese Merkmale auf und kann später prüfen, ob sie einen Einfluss auf das Zielmerkmal haben und gegebenenfalls ihren Einfluss statistisch (durch Regressionsrechnung) bereinigen.

Trotz aller Vorkehrungen werden die idealen Voraussetzungen von Bernoulli-Ketten (oder entsprechenden Verteilungsvoraussetzungen) mehr oder weniger verletzt. Patienten unterscheiden sich durch Alter, Geschlecht, Stadium der Erkrankung, allgemeine Fitness, Bereitschaft, dem Behandlungsschema zu folgen, Bereitschaft, komplementäre Heilmaßnahmen durchzuführen, soziale Einbettung in der Familie etc. Sie haben daher *keine* gemeinsame Wahrscheinlichkeit für einen Erfolg der Behandlung.

In einer frühen Phase einer Studie muss man alle Einflussgrößen analysieren. Hat man alle Maßnahmen unternommen, kann man ‚hoffen', dass die Gruppen homogen sind und dass man das Modell in

Form eines Szenarios ‚Was sagen uns die Daten, wenn wir *unterstellen*, dass die Gruppen homogen sind?' anwenden kann.

Sind die beiden Gruppen tatsächlich bezüglich einiger Confounder verschieden, so können diese die Werte des Zielmerkmals völlig überlagern. Es gab in der Geschichte der Statistik immer wieder Diskussionen, was man macht, wenn man erkennt, dass die Gruppen unterschiedlich sind und die Ergebnisse wertlos zu werden ‚drohen'. Soll man erneut randomisieren und neu zuordnen? Methodologische Gründe sprechen dagegen. Jedenfalls soll man sich über mögliche Confounder sehr viele Gedanken im Vorhinein machen und darüber Aufzeichnungen führen. Kennt man die Daten von möglichen Einflussgrößen, kann man ihren Einfluss bereinigen. Wenn man in der frühen Phase eines Projekts das übersieht, fehlen später solche Daten und man kann über die Auswirkungen nur mehr spekulieren. Flops in Anwendungen sind oft in Versäumnissen zu suchen, die in der ersten Aufarbeitung des Kontexts liegen – weit vor den eigentlichen statistischen Fehlern.

Ein Argument der Homogenisierung wird üblicherweise als Rechtfertigung für Schlüsse aus empirischen Daten herangezogen. Es fußt auf der zufälligen Zuordnung von Personen zu den Gruppen, welche miteinander verglichen werden. Weitere Untersuchungen, inwieweit das Ziel homogener und damit vergleichbarer Gruppen erreicht worden ist, werden zu selten gemacht, was die Relevanz der Ergebnisse in Frage stellt.

Ein Beispiel für die Relevanz der Überlegungen für den schulischen Unterricht

Das folgende Beispiel zeigt, dass Forschungsergebnisse unseren Alltag mitbestimmen. Ohne schulische Aufarbeitung der Methoden unter dem Gesichtspunkt der Modellierung hat man wenig Chance, die Informationen kritisch zu hinterfragen. Was soll man etwa von solchen Schlagzeilen halten? (Zeit online o.D.)

Tabelle 7: Daten zur HIV-Studie

Gruppe	HIV	Kein HIV	Alle
RV 144	51	8.160	8.201
Placebo	74	8.127	8.201
Alle	125	16.287	16.402

„Aufwind für die Aids-Forschung"

„Eine Studie belegt erstmals einen Impfschutz gegen HIV. Die Ergebnisse einer Studie in Thailand belegen erstmals, dass ein Impfstoff eine HIV-Infektion bei Erwachsenen verhindern könne.

Nach Angaben des Herstellers […] gab es mit dem Impfstoff namens RV 144 rund ein Drittel weniger HIV-Infektionen als mit einem Scheinimpfstoff (Placebo).

Thailands Gesundheitsminister […] bezeichnete das Ergebnis als "Durchbruch", da es das erste Mal war, dass ein HIV-Impfstoff vorbeugend wirkte.

Ralf Wagner, Aids-Forscher: Dieser Impfstoff ist bei Weitem das Beste, was wir je gesehen haben."

5. Zusammenfassung der Kritik

Die vorgestellte Kritik an der Abituraufgabe wird nocheinmal zusammengefasst. Die erforderlichen Modellierungsschritte scheinen tatsächlich in einer zentral gestellten schriftlichen Matura keinen Platz zu haben. Eine vollständig zentral verwaltete schriftliche Matura kann sich nur an Basiskompetenzen orientieren. Sie wird einen Sog entwickeln, der die Stochastik in Frage stellt. Dies entgegen den Erfordernissen der Praxis und vieler Studien.

Modellbildung oder vorgegebene Modelle

Die vorgestellte Aufgabe war Teil des zentralen Abiturs in Nordrhein-Westfalen. Vorgestellt wurde eine Kritik des Statistikers Davies. Es hat durchaus Kritiker gegeben (Diepgen, 2008), welche auch die Voraussetzungen der Binomialverteilung als geeignetes Modell für den Sportler Nowitzki in Frage

gestellt haben. Der Kern der öffentlichen Auseinandersetzung – Presse und der öffentliche Brief der Statistiker rund um Davies – orientierte sich an der angeblich fehlenden Angabe der Anzahl der Freiwürfe – das wurde zum Eklat ausgeweitet.

Dabei werden die einfachsten Eigenheiten eines Modells, hier der Bernoulli-Kette, glatt übersehen. Eine Bernoulli-Kette prägt einer Situation eine Struktur auf. Das geht über dieselbe Erfolgswahrscheinlichkeit und die Unabhängigkeit der Versuche hinaus. Auch Teile einer Bernoulli-Kette – so die Auswahl sich nicht an den Daten orientiert – haben grundsätzlich genau dieselben probabilistischen Eigenschaften. Jedes Spiel ist neu, jedes Spiel kann den Anfang von Beobachtungen markieren.

Daher kann man die Frage darauf abzielen, ob Nowitzki *jetzt* – damit ist genau von diesem Zeitpunkt an gemeint, an dem man ins Stadion kommt – mehr als vier Mal hintereinander einen Freiwurf verwandelt, oder ob er es nicht schafft (wir sprechen das Komplement des gesuchten Ereignisses an). Nach dem fünften verwandelten Freiwurf ist der weitere Verlauf der Versuche ohne jeden Belang. Man kommt gänzlich *ohne* die Angabe der Zahl der Beobachtungen aus. Ändert man den Wortlaut der Aufgabenstellung und führt einen Bezug zur Anzahl der Versuche ein, so wird die Aufgabe zunächst unlösbar. Gibt man eine solche Anzahl an, so legt man damit für jedes n eine neue Aufgabe fest: ‚Gelingt es Nowitzki, irgendwann im Beobachtungszeitraum der Länge n wenigstens einmal mehr als 4 Freiwürfe hintereinander zu verwandeln, oder gelingt ihm dies nicht?'

Im Wahrscheinlichkeitsteil hat sich in Lehrbüchern und Prüfungen ein Ritual eingebürgert. Eine hohle Anwendung hat eigentlich nur zum Ziel, ein paar Verteilungen auszuprobieren und Wahrscheinlichkeiten für mehr oder weniger komplexe Ereignisse zu berechnen. Ohne aufzuzeigen, dass die verwendeten Verteilungen mathematische Modelle für eine reale Situation vorzeichnen, welche der Situation implizit eine bestimmte – oft übersehene oder intuitiv schlecht durchdrungene Struktur (siehe die Bernoulli-Ketten hier) – auferlegen. Es bleibt offen, was man nach probabilistischer Modellierung der Situation besser machen kann – worin also der Vorteil besteht, solche Modelle anzuwenden; zu alledem noch, wo man weiß, dass die nötigen Voraussetzungen schlecht erfüllt sind. Für Beispiele dieser Art siehe Borovcnik (2009) oder Borovcnik und Kapadia (2011). Oder sie passen perfekt, weil man sich auf Glücksspiele alten Typs zurückzieht, die im Rückzug sind. Die neuen Glücksspiele, welche durch Programme in ‚einarmigen Banditen' implementiert sind, könnte man zu jedem Zeitpunkt steuern, sodass man ihren Mechanismus weder kennt, noch sicher sein kann, dass es einen solchen gibt.

Im statistischen Teil wurde in der geäußerten Kritik auf die Vermischung von wahren Parameterwerten und Schätzungen hingewiesen. Die Diskussion in diesem Aufsatz zeigt, dass wahre Parameterwerte immer nur relativ auf Modelle zu sehen sind und daher ihre gedankliche Einordnung als *fiktive* Werte einer statistischen Betrachtungsweise viel näher kommt. Die Beurteilende Statistik ist immer wieder mit der Unterstellung von Werten für einen Parameter konfrontiert. Man rechnet fiktive Wahrscheinlichkeiten für Fehler (1. und 2. Art) aus, um Entscheidungen zu rechtfertigen. Solche Entscheidungen spiegeln sich im Kontext durch Beantwortung von Fragen, die dem statistischen Testproblem Pate gestanden sind. Egal, wie man entscheidet, es bleibt offen, welches die wahren Werte sind.

Nach dem gängigen Bild von Stochastikunterricht hat man sich an eindeutig vorgegebenen Modellen zu orientieren und diese überhaupt nicht zu hinterfragen. Gerade die Angemessenheit von Modellen (oder besser Szenarien) ist es, was wir den Lernenden beibringen sollten. Es gibt schon viel zu viele gedankenlose Anwendungen, mit denen mehr Schaden angerichtet als Gutes getan wird. Natürlich kann man sagen, wenn ohnehin schon jetzt verabsäumt wird, Stochastik so zu unterrichten und zu prüfen, was soll sich da durch die Einführung einer zentralen Matura ändern?

Die Problematik von Teil (3) im probabilistischen Part der Nowitzki-Aufgabe zeigt auf, wie genau man den Text einer zentralen Matura eigentlich formulieren muss. Alle möglichen Missverständnisse müssen vorausgeahnt werden. Während man auf der einen Seite den Text der Aufgabenformulierung nicht sklavisch übernehmen muss (was für eine Anforderung an Maturanten!), darf man aber nicht einfach Textänderungen beliebig vornehmen. Der Aufgabenkonstrukteur muss nach Maßgabe die möglichen Missverständnisse vorausahnen und den Text entsprechend gestalten.

Schwierig wird es, wenn ein korrekter Text eine *überraschende* Lösung hat. Die Teilaufgabe (3) hatte eine solche. Das konnte aber nur deswegen „passieren", weil eine fundamentale Eigenschaft von Bernoulli-Ketten ganz allgemein im Unterricht zu wenig Beachtung findet. Eine Umformulierung einer Aufgabe muss triftigere Gründe haben als einfach „ich kann mir nicht vorstellen, dass die Aufgabe einen Sinn macht". Wenn eine Umformulierung zu kreativen Ansätzen führt, so wird man dem Prüfling einen Teil der Leistung wohl anerkennen – auch in einer zentral gestellten Prüfung.

Allerdings haben ja auch viele Schüler diese Aufgabe gelöst. Und zwar genau so, wie hier gezeigt. Diesen dann vorzuwerfen, dass sie nur aus Gedankenlosigkeit zur Lösung gelangt seien, weil ja das Beispiel eigentlich unlösbar ist, geht am Kern der Sache vorbei. Das Beispiel *ist* lösbar. Und es wirft ein durchaus bedenkliches Licht auf den Unterricht, wenn eine nicht unerhebliche Minderheit an diesem einfachen Beispiel gescheitert ist.

Der anfängliche Enthusiasmus der Aufgabenkonstrukteure wird sich – besonders nach unberechtigter oder kleinlicher Kritik – erschöpfen. Man wird sich auf Textrituale zurückziehen, die sich ausschließlich auf Basiskompetenzen beziehen. Auch Outsourcen der Aufgabenerstellung an Experten kann da nur vorübergehend helfen. Wenn es darauf ankommt, kann man, gar nicht einmal mutwillig, jeden Text zerpflücken, bis er zerfällt. So kann man an der Nowitzki-Aufgabe noch vieles mehr kritisieren, was kaum in die Diskussion eingebracht wurde – jetzt einmal abgesehen von der Problematik, ob Bernoulli-Ketten im Sport überhaupt ein adäquates Modell darstellen.

Das mag im Hinblick auf eine zentrale Aufgabenstellung in der Matura noch betrüblicher wirken. Tatsächlich kann man erwarten, dass sich gerade im Hinblick auf die beurteilende Statistik ein Kanon von rituellen Anwendungsbeispielen ergibt, der den Prozess des Modellbildens und der Beantwortung von Fragen aus dem Kontext konterkariert.

Teilzentral statt zentral organisierte Matura

Die schriftliche Mathematik-Matura hat in Österreich ab 2014 keine ‚lokalen' Anteile mehr. Die Didaktikkommission der Österreichischen Mathematischen Gesellschaft hat sich, neben vielen anderen Vereinigungen, in einer Resolution dagegen ausgesprochen (siehe ÖMG 2009); kritische Stimmen sind auch aus der Schweiz zu vernehmen (siehe VSG 2009). Man kann vorhersehen, dass sich der Unterricht auf die Vermittlung von Basiskompetenzen konzentrieren wird, also noch stärker auf das Arbeiten *im* Modell und Ausklammern von Fragen einer Modellbewertung. Das mag für viele Schüler sowieso einfacher sein. In der Folge werden im Klassengespräch und in den weiteren unterrichtlichen Bemühungen Modellbildungsprozesse mit ihren Vieldeutigkeiten wenig Rolle spielen. Auch die besseren Schüler werden damit nicht konfrontiert werden. Ohnehin stellt sich die Frage, wozu etwas lernen, wenn es nicht geprüft wird / nicht geprüft werden kann.

Dagegen sei gestellt, dass wir mehr Mathematikkenntnisse in weiteren Kreisen der Gesellschaft brauchen und mehr Kompetenzen in einschlägigen Berufen, auch in wirtschaftswissenschaftlichen Bereichen. Die Top 5% der Schüler kann wohl kein Schulsystem ‚gefährden'. Wir bräuchten aber – vor allem in technischen Studien – mehr an Vorkenntnissen in Mathematik. Insgesamt ist eine breitere

Basis bei den Top 25-30% vordringlich. Es sei der Spekulation der Leser überlassen und der zukünftigen Entwicklung, ob wir das mit den nun gesetzten Rahmenbedingungen erreichen. Der Vorschlag des Autors war immer: Eine *teil*zentrale schriftliche Matura. Die zentral gestellten Aufgaben dienen dabei zur Überprüfung einer gemeinsamen Sprache und von Basiskompetenzen. Lokal gestellte Aufgaben lassen den Lehrern überdies einen gewissen Freiraum zum eigenen Engagement, das etwa zum sinnvollen Anwenden und Hinterfragen von Ergebnissen genützt werden kann.

Der Autor möchte an dieser Stelle Herrn Hans Humenberger für die umsichtige und kritische Diskussion des Artikels herzlich danken. Seine Anmerkungen haben wesentlich zur Verbesserung der Darlegung der Gedanken beigetragen.

Literatur

Borovcnik, M. (2009): Aufgaben in der Stochastik – Chancen jenseits von Motivation. *Schriftenreihe Didaktik der Mathematik der Österr. Math. Ges.* 42, 20-42.

Borovcnik, M. (2009): ‚Anwendungen' und Anwendungen – Zentrales Abitur und vergebene Chancen für den Unterricht in Stochastik. *Stochastik in der Schule* 29 (3), 9-18.

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

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

Davies, P. L. (2009): Einige grundsätzliche Überlegungen zu zwei Abituraufgaben. *Stochastik in der Schule,* 29 (2), 2-7.

Davies, L., Dette, H., Diepenbrock, F.R., & Krämer, W. (2008). Ministerium bei der Erstellung von Mathe-Aufgaben im Zentralabitur überfordert? *Bildungsklick*. Online: http://bildungsklick.de/a/61216/ministerium-bei-der-erstellung-von-mathe-aufgaben-im-zentralabitur-ueberfordert/ (Einsicht: 25.10.2010).

Diepgen, R. (2008): Kein Witz!? Zur Nowitzki-Aufgabe im NRW-Zentralabitur 2008. *Stochastik in der Schule* 28 (3), 20-28.

Eichler, A.; & Vogel, M. (2010). Daten und Zufall als einende Leitidee. *Aufsatz präsentiert an der DAGStat2010*, Dortmund, 25.03.2010.

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

Mosler, K. & Schmid, F. (2006): *Wahrscheinlichkeitsrechnung und schließende Statistik*. Berlin: Springer.

ÖMG (2009). *Stellungnahme zur Zentralmatura in Österreich*. Online: http://www.oemg.ac.at/DK/index.html (Einsicht: 25.10.2010).

Schulministerium NRW (o.D.). Zentralabitur NRW. Standardsicherung. Online: www.standardsicherung.nrw.de/abitur-gost/fach.php?fach=2. Direkter Link zur Prüfungsaufgabe: http://www.standardsicherung.nrw.de/abitur-gost/getfile.php?file=1800 (Einsicht: 25.10.2010).

VSG (2009): Die Zukunft des Gymnasiums – Positionspapier des VSG. *Verein Schweizerischer Gymnasiallehrerinnen und Gymnasiallehrer*. Online: http://www.vsg-sspes.ch/fileadmin/files/pdf/09.03_Zukunft_Gymnasium_details_d.pdf (Einsicht: 25.10.2010).

Zeit online (o.D.): *Aufwind für die Aids-Forschung*. Online: http://www.zeit.de/wissen/gesundheit/2009-09/aids-studie-thailand (Einsicht: 25.10.2010).