# „Anwendungen" und Anwendungen – Zentrales Abitur und vergebene Chancen für den Unterricht in Stochastik

MANFRED BOROVCNIK, KLAGENFURT

***Zusammenfassung.*** *In Nordrhein-Westfalen wurde die ‚Nowitzki-Aufgabe' in ‚unlösbarer' Form gestellt. Das hat zu einem erheblichen Aufruhr und einer erbitterten Diskussion um das Zentralabitur geführt. Natürlich ist es mehr als peinlich, wenn solch ein Fehler unterläuft. Allerdings kann die geäußerte Kritik nicht in allen Teilen einer näheren Inspektion standhalten. Wie genau man auch formuliert, fast jede Aufgabe kann man mehrdeutig auffassen. Das hat für zentral gestellte Aufgaben weitreichende Folgen, mehr als für lokale. Darüber hinaus wurde aber kaum in die Diskussion eingebracht, dass die verwendete Modellierung überhaupt nicht angemessen ist und damit die gestellten Fragen nicht sinnvoll beantworten lässt.*

## Vorausblick

Die Nowitzki-Aufgabe wurde an vielen Stellen kritisiert. In diesem Aufsatz geben wir den Kern der Kritik von Davies (2009) wieder. Es soll klar werden, dass die Kritik nicht in allen Punkten einer Kritik standhält. Dabei wird auf eine fundamentale Eigenschaft des verwendeten Modells einer Bernoulli-Kette eingegangen.

Das Modell einer Bernoulli-Kette birgt weiters in sich Eigenschaften, die es für eine Anwendung im Sport als völlig ungeeignet erscheinen lassen. Das Verhältnis zwischen Modell und der damit modellierten realen Situation muss geeignet reflektiert werden. Solche Aspekte aus dem Abitur gänzlich wegzulassen bedeutet, wesentliche Chancen eines anwendungsorientierten Stochastikunterrichts zu vergeben.

## Eine zentral gestellte Abituraufgabe, die Staub aufgewirbelt hat

Der deutsche Basketball-Profi Dirk Nowitzki spielt in der amerikanischen Profiliga [...]. In der Saison 2006/2007 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 Freiwürfen 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 [...] 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.

## Worum es geht

### Wahrscheinlichkeiten und Schätzwerte

Davies schreibt: „Wirft man eine 1-€-Münze, beträgt die Wahrscheinlichkeit für einen Zahlwurf 1/2. Würfelt man mit einem normalen Würfel, beträgt [...] für eine bestimmte Augenzahl 1/6. Beim Lotto beträgt [...] 1/49.

Diese Wahrscheinlichkeiten werden durch Symmetrieüberlegungen ermittelt. [...], wo es keine Symmetrieargumente gibt, z. B. die Erfolgswahrscheinlichkeit bei Freiwürfen im Basketball, greift man auf empirische Ergebnisse zurück und die Wahrscheinlichkeit wird *geschätzt*."

Davies führt dazu die Mädchenquoten in drei Krankenhäusern an: „Die 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."

| Krankenhaus | Geburten | Mädchenquote |
|---|---|---|
| A | 514 | 0,492 |
| B | 358 | 0,450 |
| C | | 0,508 |

**Anm. 1:** Natürlich könnten die drei Krankenhäuser echt verschiedene Mädchenquoten haben. Es lassen sich durchaus Störfaktoren denken, wie etwa ein Privatarzt mit Vertrag mit dem Haus C betreue nur Mädchen-Schwangerschaften.

„In der mathematischen Statistik unterscheidet man zwischen einer wahren Wahrscheinlichkeit und einer geschätzten Wahrscheinlichkeit. Diese begriffliche Unterscheidung ist von grundlegender Bedeutung und unerlässlich. [...] die beiden Begriffe vermischt [...]"

**Anm. 2:** Mein akademischer Lehrer Walther Eberl würde auch entgegnen: ‚Es gibt keinen wahren Wert, der hängt vom Messverfahren ab.' Positivisten haben mit der Gleichsetzung von Wirklichkeit und Modell durchaus Schwierigkeiten; die Gleichsetzung von ‚wahrem' Wert und Schätzwert wird – zu Recht – kritisiert; dagegen wird ‚übersehen', dass ein wahrer Wert von der Betrachtungsweise, also von Hypothesen, d. h. von der genauen Festlegung des Modells, abhängen kann.

So etwa kann man fragen, ob die Trefferquote für Heimspiele mit jener für Auswärtsspiele übereinstimmt, oder ob es da Unterschiede gibt. In der beurteilenden Statistik fragt man oft, ob die verfügbaren Daten mit einem aus Hypothesen abgeleiteten Wert eines Parameters vereinbar sind. Die Sprechweise „wahrer Wert" engt dieses Denken in unterstellten Werten unnötig ein. Damit man die Nowitzki-Aufgabe lösen kann, muss man für die Trefferquote Annahmen unterstellen, konkrete Werte aus Daten schätzen oder aus Hypothesen einsetzen etc.

Davies setzt fort: „Somit ist die Aufgabe schlecht gestellt und die Lösung des Ministeriums falsch."

**Anm. 3:** Die Aufgabe an sich hat mit der von Davies angesprochenen fehlenden begrifflichen Unterscheidung nichts zu tun; daher ist die Feststellung, sie sei „schlecht gestellt", vorschnell. Gegen die Aufgabe selbst richtet sich die angesprochene Kritik im Weiteren auch – leider – nicht, sieht man von der als fehlend bemängelten Angabe in (3) von Teil a) einmal ab.

## Modellierung

Bei der Aufgabe handelt es sich um eine Anwendung der Binomialverteilung; mit dieser modelliert man Vorgänge,

- bei denen eine feste Anzahl $n$ von Versuchen durchgeführt wird,
- bei jedem Versuch gibt es nur zwei Möglichkeiten (Erfolg, Misserfolg).

„Es wird davon ausgegangen, dass [wie üblich]

- die Erfolgswahrscheinlichkeit $p$ für alle Versuche dieselbe ist und
- dass sich die einzelnen Versuche nicht gegenseitig beeinflussen."

Die Voraussetzung für diese Art von ‚Stichproben' nennt man Bernoulli-Kette. Ergebnisse der einzelnen Versuche werden durch Zufallsgrößen $X_1, ..., X_n$ bezeichnet:

$$X_i = 1 \text{ (Erfolg) bzw. } X_i = 0 \text{ (Misserfolg).}$$

„Bei einer Textaufgabe muss die Modellierung entweder angegeben werden, oder sie muss aus der Beschreibung der Situation eindeutig hervorgehen.

Wenn die Modellierung feststeht, wird die Aufgabe innerhalb des Modells weiter bearbeitet. D. h., es wird angenommen, dass das Modell stimmt. Es ist nicht die Aufgabe des Abiturienten, die Angemessenheit der Modellierung zu hinterfragen."

**Anm. 4:** Auch unter dem Aspekt einer fair gestellten Prüfung wird man ein gewisses Maß an Modellkritik von einem Abiturienten erwarten können.

Bei der Binomialverteilung ist die Anzahl $n$ von Versuchen bekannt. Bei der Erfolgswahrscheinlichkeit $p$ gibt es zwei Möglichkeiten:

- $p$ ist bekannt oder
- $p$ ist nicht bekannt und man will $p$ aus der Versuchsreihe schätzen.

„In der letzteren Situation muss klar zwischen dem wahren Wert $p$ und einem Schätzwert $\hat{p}_{\mathrm{n}}$ unterschieden werden. [...] Die klare Unterscheidung zwischen einem wahren Parameterwert und einem Schätzwert hierfür ist fundamental: Die ganze schließende Statistik basiert darauf."

## $p$ bekannt

Woher kennt man nun $p$? Bei einer normalen Münze scheint 1/2 aus Symmetriegründen plausibel, denn „es gibt im Normalfall keinen Grund, die eine oder die andere Seite der Münze vorzuziehen. [...] Um ganz eindeutig zu sein, [...] ‚fairen' oder ‚unverfälschten' Münze sprechen."

Davies schreibt weiter:

- „Die Wahrscheinlichkeit $p$ ist [bei Münzen, Würfeln etc.] durch Symmetrieüberlegungen bestimmt."
- „In den meisten Anwendungsfällen ist aber eine Symmetrie nicht vorhanden und […] $p$ muss anders ermittelt werden."

Man kann schreiben

- „Aus langjähriger Erfahrung [...] Mädchen [...] 0,481 [...]"
- „bei einem gefälschten Würfel [...] für die Augenzahl Sechs 0,183".

„Es ist nicht Aufgabe des Abiturenten, nachzufragen, wie die Werte 0,481 oder 0,183 ermittelt wurden, er hat sie lediglich bei den weiteren Berechnungen einzusetzen. [...] Auf jeden Fall gibt es bei bekanntem $p$ eine klare Sprachregelung, die einzuhalten ist [...].“

**Anm. 5:** Oft ist die Annahme einer Gleichwahrscheinlichkeit mehr eine *Sprechweise*, um ein Modell zu benennen. Mangels genauerer Daten über *diese* Münze und *diese* Art des Werfens ist das ein brauchbares Modell. Man rechnet in der Praxis häufig mit *fiktiven* Werten für Parameter, ganz einfach, um die Konsequenzen daraus zu bestimmen. Wahrscheinlichkeiten werden hierbei mehr auf der Basis von Szenarien (was wäre, wenn ...?) verwendet als auf der Basis von Modellen, die ziemlich genau auf die reale Situation passen sollten.

Die Annahmen einer Bernoulli-Kette passen denn auch nicht für Anwendungen im Sport. Dennoch könnte man eine Bernoulli-Kette einfach als *Szenario* auffassen, dessen Folgerungen einer theoretischen Analyse unterworfen werden. Dann muss man allerdings auch noch Situationen finden, für die dieses Szenario relevante Ergebnisse liefern kann (siehe Anmerkung 8).

### $p$ unbekannt

In diesem Fall muss man die unbekannte Wahrscheinlichkeit aus einer Stichprobe schätzen, mit all den Voraussetzungen, die man an eine Stichprobe stellen muss (und die in der Praxis eigentlich nicht erfüllbar sind): dieselbe Erfolgswahrscheinlichkeit bei jeder Messung, Unabhängigkeit der Versuche untereinander. Davies beschreibt die Situation so:

„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.“

Die Ungenauigkeit von Schätzwerten beurteilt man normalerweise aufgrund von Konfidenzintervallen; das illustriert Davies anhand von unterschiedlichen Stichprobenumfängen.

**Anm. 6:** Natürlich spricht Davies davon, dass man auch nach 1 Mio. Daten nur einen „Schätzwert“ und nicht die genaue Wahrscheinlichkeit kennt. Viel wichtiger als so viele Daten zu haben ist die Qualität der erhaltenen Daten: Diese sinkt meist mit der absoluten Zahl der Daten rapide, sodass daraus berechnete Konfidenzintervalle wenig aussagen. Die Datenbasis zu hinterfragen ist Haupttätigkeit des praktischen Statistikers, die Intervalle berechnet man bequem mit Software.

## Teil a) der Aufgabe

### Teile (1) und (2)

„Interpretiert man 90,4 % als eine Wahrscheinlichkeit von 0,904, so handelt es sich eindeutig um einen Schätzwert [...] und nicht um die wahre unbekannte Wahrscheinlichkeit $p$. [Dagegen] wird die Quote 0,904 vom Ministerium als die wahre Wahrscheinlichkeit betrachtet:

‚Die Zufallsvariable $X$ für die Anzahl der Treffer bei 10 Versuchen ist B(10; 0,904)-verteilt.‘

[...] Es liegt also eine Begriffsverwirrung vor, weil nicht unterschieden wird zwischen der wahren Wahrscheinlichkeit $p$ und einem Schätzwert $\overline{X}_n$ für diese Wahrscheinlichkeit.

[...] In Teil b der Aufgabe [...] erzielte Nowitzki 498 Treffer bei 551 Versuchen. Berechnet man hierfür ein 95%-Konfidenzintervall für die wahre Wahrscheinlichkeit $p$, so bekommt man

[0,8792, 0,9284].

Damit ist die einzig plausible Lösung zu (1) [...], dass die gesuchte Wahrscheinlichkeit zwischen [diesen beiden Werten] liegt.“

**Anm. 7:** Die Saison 2006/2007 ist gelaufen. Man kennt *alle* Daten. Es wird nicht mehr davon geben. Eine neue Saison wird unwiederbringlich andere Konstellationen haben. Die Erfolgsrate in dieser Saison beträgt (als Faktum) 0,904. Man könnte noch hinterfragen, ob es einen Sinn macht, diese Zahl als Wahrscheinlichkeit zu interpretieren, weil das ja voraussetzt, dass in jedem Freiwurf für Nowitzki genau dieselben Bedingungen geherrscht haben, unabhängig voneinander und unabhängig vom Spielverlauf, Spielstand etc. So gewendet könnte man fragen, ob die Daten tatsächlich mit den Voraussetzungen einer Bernoulli-Kette verträglich sind. Man könnte also fragen, ob die Zahl der ‚Runs‘ (der Erfolge oder der Misserfolge hintereinander) über oder unter dem erwarteten Maß für eine Bernoulli-Kette liegt oder nicht.

„Wenn der Stichprobenumfang, auf dem der Schätzwert basiert, hinreichend groß ist, ist die Differenz klein, und gerade diese Tatsache rechtfertigt Formulierungen wie ‚aufgrund langjähriger Erfahrung weiß man, dass die Erfolgswahrscheinlichkeit 0,9 beträgt‘. Eine solche Formulierung wurde nicht gewählt, im Gegenteil, der Hinweis auf die Saison 2006/2007 verdeutlicht, dass der Stichprobenumfang eher klein war.“

**Anm. 8:** Der Umfang der Stichprobe hilft hier gar nichts, denn mit dem Mischen von Saisonen wird eine Erfolgswahrscheinlichkeit immer fragwürdiger.

Je länger die Serien, desto weniger werden die Voraussetzungen einer Bernoulli-Kette plausibel.

Wenn relative Häufigkeiten als Schätzwerte für zugrunde liegende Wahrscheinlichkeiten herangezogen werden sollen, müssen die Bedingungen einer Bernoulli-Kette für die Entstehung der Daten gelten. Erst dann werden Schätzwerte aus größeren Stichproben auch genauer.

Allerdings muss man für Nowitzkis Treffer die Voraussetzungen ziemlich ‚strecken': Nehmen wir den Wert 0,904 als Schätzwert für die gegenwärtige Kompetenz – formalisiert als Wahrscheinlichkeit $p$ – einen Freiwurf zu verwandeln. Dahinter steckt die unrealistische Annahme einer ‚Homogenisierung': Nowitzkis Fähigkeit war die ganze Saison konstant und unabhängig von allen begleitenden Umständen.

Für die *Gesamtheit* der Freiwürfe mag die Homogenisierung eine brauchbare Annahme sein: Vielleicht mitteln sich die Abweichungen von den Annahmen in einer längeren Serie irgendwie aus. Die Ergebnisse jedoch für irgendwelche 10 Würfe anzuwenden und die Wahrscheinlichkeit, höchstens 8 Treffer bei 10 Versuchen daraus zu berechnen, macht wenig Sinn. Schon gar nicht, wenn es die letzten 10 eines entscheidenden Spiels sind.

Wir müssten wohl merkwürdige Szenarien wie das folgende heranziehen, damit die Ergebnisse der Binomialverteilung umsetzbar werden: Alle Freiwürfe der gesamten Saison werden per Videokamera aufgezeichnet. Jetzt wählen wir 10 Aufzeichnungen per Zufall aus: Wie oft hat Nowitzki getroffen?

### Teil (3)

„[...] lässt mehrere Interpretationen zu."

- „Nimmt man an, dass 10 Versuche durchgeführt wurden, ist die Aufgabe nun wohl gestellt, aber die Berechnung der Lösung ist äußerst langwierig. Von den $2^{10} = 1024$ möglichen Versuchsreihen muss man die Anzahl [...] bestimmen, in denen die Eins höchstens viermal nacheinander vorkommt. [...] Es ist klar, [...] Schwierigkeitsgrad und [...] Zeitaufwand [...] nicht angemessen sind. "

Davies rechtfertigt diese Variante – wohl aus der Sicht der zu Prüfenden – so: „Da aber die Anzahl 10 in (1) und (2) angegeben wird, liegt es nahe, auch 10 in (3) anzunehmen."

**Anm. 9:** In den hier ausgelassenen kombinatorischen Überlegungen könnte man durch strukturiertes Abzählen der Fälle den Aufwand erheblich verkürzen. Letztlich muss man aber die Ansicht von Davies teilen, wenngleich die Begründung für 10 Versuche trotz einer gewissen ‚Verbreitung' einer inneren Logik entbehrt.

- „Wir nehmen nun an, dass die Anzahl der Versuche $n = 5$ ist.
  Das Einzige, das für diese Interpretation spricht, ist, dass die Lösung mit der Lösung des Ministeriums übereinstimmt. Zu der Lösung schreibt das Ministerium:
  ‚Man betrachtet das Gegenereignis, dass er fünf Treffer hintereinander schafft ...'
  Das Gegenereignis von ‚höchstens vier' ist nicht ‚fünf', sondern ‚mindestens fünf'. **Es ist nur dann fünf, wenn er genau fünfmal wirft".**

Davies schreibt,
„es gibt keinen Grund, dies anzunehmen."

**Anm. 10:** Ja, es gibt keinen Grund anzunehmen, dass es fünf Versuche gibt, wieso auch!

- „[...] stellen wir uns vor, dass beim Training folgendes Spiel zwischen zwei Spielern [...] Einer fängt an [...] und macht Freiwürfe bis zum ersten Fehlwurf. Der zweite Spieler ist nun an der Reihe [...] bis zum ersten Fehlwurf usw. Nowitzki spielt dieses Spiel gegen einen Teamkameraden und fängt an. [...], dass er bei seinem ersten Versuch höchstens viermal erfolgreich war."

Die möglichen Versuchsfolgen sind

0 10 110 1110 11110

„Man stellt fest, dass [die Lösung 0,3963] 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 [...]"

**Anm. 11:** Der Versuch, die Aufgabe zu retten, ohne die ‚fehlende' Angabe – durch Raten etwa – zu ergänzen, bringt tatsächlich die Lösung. Allerdings ist die dargestellte Situation künstlich und unnötig kompliziert. Tatsächlich ist die Aufgabe *ohne* die Anzahl der Freiwürfe lösbar, auch *ohne* auf ein solches Szenario ausweichen zu müssen; man muss nur grundsätzliche Eigenschaften einer Bernoulli-Kette ausnützen. Wir werden die Lösung weiter unten darstellen und diese wichtige Eigenschaft von Bernoulli-Ketten wiederholen oder vielleicht erst ins rechte Licht rücken.

## Teil b) der Aufgabe

„Führt man $n$ Versuche durch mit einer konstanten Erfolgswahrscheinlichkeit, so ist der Erwartungswert der Trefferanzahl

$$T_n = X_1 + ... + X_n{:}\ \ E(T_n) = n\,p.$$
[...] Dabei ist $p$ der wahre unbekannte Wert für die Wahrscheinlichkeit."

### Ein „Tippfehler": Trefferzahl statt Trefferquote

„[dass wir] zwischen Heim- und Auswärtsspielen unterscheiden müssen."

Erfolgswahrscheinlichkeit für

Heimspiele: $p_H$ – Auswärtsspiele: $p_A$.

Dann ist der Erwartungswert der Trefferanzahl

Heimspiele: $E(TH_{n_H}) = n_H \cdot p_H$
Auswärtsspiele: $E(TA_{n_A}) = n_A \cdot p_A$.

„Nun, die wahren Werte [für die Trefferwahrscheinlichkeiten] kennen wir nicht. Das Ministerium ersetzt sie einfach durch die Schätzwerte [...]

‚Es sei $X$ B(263; 0,904)-verteilt',

woraus die Verwechslung eines Schätzwertes mit dem wahren Parameterwert sichtbar ist.

Damit ist der ‚Erwartungswert' für Heim- und Auswärtsspiele nun nichts anderes als

$$TH_{n_H} + TA_{n_A}$$

Die Trefferanzahl bei Auswärtsspielen ist $TA_{n_A}$ und wir müssen nun testen, ob [diese Zahl] signifikant kleiner ist als $TH_{n_H} + TA_{n_A}$.

Dies ist nun ja immer der Fall [...]."

So schließt Davies diesen Punkt ab und setzt fort:

„Setzt man die Daten ein, müssen wir testen, ob
$231 < 267 + 231$, was ja stimmt."

Jetzt erklärt Davies den ‚offensichtlichen' Fehler so:

„[...] vermutlich ein Tippfehler vor: [...] nicht die ‚Trefferanzahl', sondern [...] Trefferquote."

**Anm. 12:** Man kann nicht einfach zwei ‚Zahlen' statistisch miteinander vergleichen, also fragen, ob 231 signifikant kleiner als 267 + 231 ist. Man hat den empirisch beobachteten Wert einer Zufallsvariable mit einer hypothetisch angenommenen Verteilung für diese Zufallsvariable zu vergleichen.

Natürlich ist die gestellte Frage mit Trefferanzahlen lösbar – Trefferanzahlen und Trefferquoten sind ja bis auf Normierung gleichwertig; ist das Problem also mit Trefferquoten lösbar, so auch mit Trefferanzahlen. Wir zeigen die Lösung im Detail, weil sie die Denkweise der Umsetzung von Sachfragen in statistische Hypothesen und deren statistische Überprüfung aufzeigt.

### Lösung von Teil (1)

Betrachten wir die drei Bernoulli-Ketten von Heim- und Auswärtsspielen sowie allen Spielen. Nach Ende der Saison ‚gibt' es faktisch die Wahrscheinlichkeiten für den Erfolg sowie alle Trefferzahlen.

| | Treffer | Versuche | Ws. faktisch |
|---|---|---|---|
| heim | $TH = 267$ | $n_H = 288$ | $p_H = \frac{267}{288} = 0{,}927$ |
| auswärts | $TA = 231$ | $n_A = 263$ | $p_A = \frac{231}{263} = 0{,}878$ |
| alle | $Tg = 498$ | $n_g = 551$ | $p_g = \frac{498}{551} = 0{,}904$ |

Man betrachtet nun alle Auswärtsspiele und fragt, ob die Trefferzahl signifikant über der für Heim- und Auswärtsspiele liegt. Gibt man alle Spiele in einen Topf, so erhält man $p_g = 0{,}904$. Man vergleicht, ob in den $n_A = 263$ Auswärtsversuchen die Zahl der beobachteten Treffer mit $TA = 231$ signifikant unter der einer Bernoulli-Kette mit Erfolgswahrscheinlichkeit 0,904 zu liegen kommt. Erwartungswert dieser Kette von Spielen ist *nicht* 267 + 231, wie Davies angibt, sondern

$$n_A\,p_g = 263 \cdot 0{,}904 = 237{,}7.$$

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

Wahl des statistischen Modells

$$TA_{n_A} \sim \mathrm{B}(n_A, \pi),$$

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

Wahl der Hypothesen

Als Nullhypothese wählen wir
$$\pi = \pi_g,$$
wobei $\pi_g$ die Trefferstärke in allen Spielen darstellt. Als Alternativhypothese könnten wir die einseitige Hypothese
$$\pi < \pi_g$$
wählen (im Sport allgemein macht der Heimvorteil etwas aus). Es macht einen Sinn, $\pi_g$ durch den Schätzwert 0,904 zu ersetzen; andere Information über alle Spiele haben wir nicht. (Wir könnten auch mit ‚ungünstigsten' Werten für alle Spiele der Saison arbeiten, um zu berücksichtigen, dass wir nur über einen Schätzwert verfügen; das tut man in aller Regel nicht, wenngleich es manches Mal angezeigt wäre,

weil die vorhandene Information sehr ungenau ist. Im vorliegenden Beispiel ist sie relativ präzise.)

Bestimmung des Ablehnbereichs

Jetzt bestimmen wir die kritische Grenze für einen Test mit 5 % Fehler 1. Art mit $c_{5\%} = 229$. Das bedeutet, die Nullhypothese ist abzulehnen, falls $TA_{n_A} \leq 229$ ist.

Entscheidung aufgrund der beobachteten Daten

Weil $TA_{n_A} = 231$ beobachtet wurde, kann man die Nullhypothese nicht ablehnen. Fazit: Es kann statistisch (auf dem Signifikanzniveau 5 %) nicht nachgewiesen werden, dass Nowitzki in Auswärtsspielen schlechter ist als in der *gesamten* Saison.

Tatsächlich ist der p-Wert (p value, hat nichts mit dem Parameter $p$ zu tun) der Beobachtung 231 gleich 0,1001. D. h., eine solche Beobachtung (einseitig getestet) und noch kleinere Erfolgszahlen hat immerhin eine Wahrscheinlichkeit von knapp mehr als 10 % – daher ist *keine* signifikante Abweichung der Trefferzahl bei Auswärtsspielen im Vergleich zu allen Spielen festzustellen.

**Anm. 13:** In der Beurteilung, ob Nowitzki in Auswärtsspielen schlechter ist als in Heimspielen, ist das *Szenario einer Bernoulli-Kette* viel aussagekräftiger als bei der Berechnung von einzelnen Wahrscheinlichkeiten. Tatsächlich werden ganze Blöcke von Spielen gegeneinander verglichen 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 die statistische Beurteilung der anstehenden Sachfrage kann daher das Szenario einer Bernoulli-Kette, obwohl es nicht wirklich besonders gut passt, relevante Ergebnisse ermöglichen – dies in Abgrenzung zum Wahrscheinlichkeitsteil der Aufgabe, wo wir uns vergeblich bemüht haben, Situationen aufzuzeigen, in denen dieses Szenario anwendbar sein könnte.

**Anm. 14:** Die Frage (1) von Aufgabenteil b) ist ungeschickt gestellt, denn alle Spiele beinhalten auch die Auswärtsspiele; die kleinere Erfolgsrate auswärts senkt ja auch die aller Spiele. Man hat eigentlich Auswärts- gegen Heimspiele zu vergleichen, wie das in (2) angestrebt ist. Die Vorgangsweise bei der Lösung wird im Folgenden knapp wiedergegeben.

## Lösung von Teil (2)

Hier sind die Auswärts- mit den Heimspielen hinsichtlich der Trefferzahl zu vergleichen. Für Heimspiele hat man eine Erfolgswahrscheinlichkeit von $p_H = 0{,}927$. Man fragt, ob die beobachtete Anzahl von $TA = 231$ in Auswärtsspielen für eine Bernoulli-Kette der Länge $n_A = 263$ mit $0{,}927$ signifikant zu klein ist. Hier wären 243,8 Erfolge zu erwarten; der p-Wert der Beobachtung liegt nun bei 0,0032! Auswärtsspiele unterscheiden sich demnach hochsignifikant von Heimspielen.

**Anm. 15:** Ohne Schätzung der Erfolgswahrscheinlichkeiten bei Heim- und Auswärtsspielen könnte man die Frage nach den Unterschieden mit dem $\chi^2$-Unabhängigkeitstest beantworten. Es ergibt sich $\chi^2 = 3{,}76$ bei einem p-Wert von 0,0525.

| Trefferzahlen | beobachtet | | | erwartet bei Unabhängigkeit | | |
|---|---|---|---|---|---|---|
| | Treffer | Niete | Versuche | Treffer | Niete | Versuche |
| heim | 267 | 21 | 288 | 260,3 | 27,7 | 288 |
| auswärts | 231 | 32 | 263 | 237,7 | 25,3 | 263 |
| alle | 498 | 53 | 551 | 498 | 53 | 551 |

$\chi^2 = 3{,}76$, p-Wert = 0,0525

Weniger Information verschenkt man, wenn man auf den exakten Test von Fisher zurückgreift. Dabei fragt man – im obigen Tableau – gemäß der Sachfrage Folgendes:

Alle 551 Versuche sind als Kugeln in einer Urne repräsentiert; 498 weiße Kugeln stellen die Treffer, 53 schwarze Kugeln die Nieten dar. Man hat nun für die Versuche in Auswärtsspielen 263 Kugeln gezogen und dabei ‚nur' 231 weiße erhalten.

> Wie wahrscheinlich sind 231 oder gar weniger weiße Kugeln?

Die Antwort erhält man aus der hypergeometrischen Verteilung. Die Beobachtung von 231 hat einen p-Wert von 3,6 %. Es ist daher auf dem 5 %-Signifikanzniveau statistisch gesichert, dass Nowitzki auswärts schwächer ist als daheim.

Man beachte: Es wurden hierbei weder die Spielstärken in Heim- oder Auswärtsspielen noch 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.

### Jetzt mit der Trefferquote

Für die mittlere Trefferquote für Heimspiele sei statt der Bezeichnung in Davies folgende Notation verwendet: $\overline{TH}_{n_H}$; entsprechend für Auswärtsspiele: $\overline{TA}_{n_A}$. Für die Erwartungswerte der Trefferquoten für Heim- bzw. Auswärtsspiele hat man

$$E(\overline{TH}_{n_H}) = p_H \text{ und } E(\overline{TA}_{n_A}) = p_A$$

Davies führt aus: „Der Erwartungswert der Trefferquote für Heim- und Auswärtsspiele beträgt

$$[p_g =] \frac{n_H \cdot p_H + n_A \cdot p_A}{n_H + n_A}$$

und wir sollen nun anhand der Daten testen, ob

$$p_A < [p_g]$$

Da aber die wahren $p_H$ und $p_A$ nicht bekannt sind, hat das Ministerium sie [...] einfach durch [deren Schätzwerte] ersetzt und tat so, als ob sie die wahren Werte sind [...]. Gleichzeitig wird aber in der Lösung des Ministeriums [die beobachtete Erfolgsrate in Auswärtsspielen] als Schätzwert für $p_A$ behandelt. [...] wird manchmal als der wahre Wert betrachtet und manchmal als ein Schätzer hierfür und noch dazu im selben Satz. Die Begriffsverwirrung ist komplett.“

**Anm. 16:** Die Überlegungen zur Trefferzahl sind auch hier anwendbar. Es ist dabei egal, ob man nach der Saison die verfügbaren Daten als Einheit sieht (und $p_g = 0{,}904$ bestimmt) oder ob man eine fiktive größere Gesamtheit ‚aller‘ Spiele annimmt (und $\hat{p}_g = 0{,}904$ schätzt, aber mit diesem Schätzwert ein Modell für die Bernoulli-Kette spezifiziert – in diesem Fall müsste man allenfalls noch mit besten und schlechtesten Werten, die man aus einem Konfidenzintervall erhält, die Genauigkeit des Ergebnisses evaluieren; das unterbleibt in der Praxis – leider – meistens).

Man fragt wieder, ob die empirischen Daten für die Erfolgsrate für Auswärtsspiele (die zwar numerisch identisch mit der Schätzung oder dem ‚wahren‘ Wert $p_A$, aber nicht als Schätzwert dafür verwendet wird) mit dem unterstellten Modellwert

$$p = p_g = 0{,}904$$

vereinbar ist oder ob man die Abweichungen als signifikant einstufen kann.

Jeder Wert kann unter verschiedenen Gesichtspunkten einmal als Schätzwert für eine unbekannte Wahrscheinlichkeit oder als Realisierung einer Zufallsvariablen betrachtet werden. Man muss nur immer wissen, was wann erfolgt.

**Anm. 17:** Das Beispiel hat auch einen Vorzug: Die Umsetzung von der Fragestellung in statistische Hypothesen ist wirklich zu leisten. Die ‚Ableitung‘ der Formulierung statistischer Hypothesen aus dem Kontext, also die Umsetzung von Fragestellungen aus dem Sachkontext in die Sprache der Statistik, ist meist vorweggenommen. Es wird einfach hingestellt, dass zu testen ist, ob der Parameter etwa $\pi = 0{,}5$ oder irgendein anderer Wert ist. Über die Festlegung von Parametern, Wahl der Testgröße oder Alternativen gibt es keinerlei Zweifel. Diese sind vorgegeben. Im Sinne von Modellierung stellen übliche Beispiele keinerlei Anforderungen.

**Anm. 18:** Man kann das Beispiel auch mit einem Test für die *Differenz von zwei Anteilen* lösen. Dabei werden beide Anteile, sowohl die für Heim- als auch für Auswärtsspiele, geschätzt und mit der beobachteten Differenz dieser Anteile wird dann getestet, ob $\pi_A - \pi_H = 0$ oder ob $\pi_A - \pi_H < 0$ ist. Wenn man die Frage statistisch so umsetzt, muss man die Verteilung der Differenz von Anteilen kennen und wissen, wann diese durch eine Normalverteilung approximiert werden kann. Wenn man statt Anteilen Mittelwerte auf Differenz prüft, landet man beim sogenannten Welch-Test.

Beide Situationen sind selbst für die einführende Vorlesung an Universitäten zu hoch gegriffen, wenngleich Anwender mit solchen Fragestellungen oft konfrontiert sind. Für die Darstellung der Vorgehensweise und die nötigen Approximationen der Testverteilung sei auf die gut lesbaren Bücher von Mosler & Schmid (2006) oder Lorenz (1996) verwiesen. Die Ausführungen in diesem Aufsatz zeigen, dass die Sachfrage *ohne* diese alternative Modellierung zufriedenstellend gelöst werden kann und häufig auch so gelöst werden soll.

## Umgehen mit den Voraussetzungen

### Bernoulli-Ketten als Modell im Sport

Wer auch nur ein bisschen Ahnung von Sport hat, wird die Voraussetzungen einer Bernoulli-Kette weit von sich weisen:

- Bei jeder Durchführung hat man dieselbe Erfolgswahrscheinlichkeit.
- Die Wahrscheinlichkeit für Erfolg beim i-ten Versuch ist unabhängig von vorherigen Ergebnissen.

Erfolgswahrscheinlichkeiten im Sport sind nämlich

- während einer Saison starken Schwankungen unterworfen; Heim- und Auswärtsspiele sind nur ein weiterer Faktor;
- stark vom Verlauf des Spielgeschehens **abhängig**

– und natürlich auch von den vorherigen Ergebnissen. Man spricht von einem ‚Lauf' oder auch von einer ‚Pechserie'.

Damit wird die Aufgabe zu einer entarteten „Anwendung“: Wir setzen apodiktisch fest, dass die Voraussetzungen zu gelten haben. Dann rechnen wir einfach damit.

### Lösung von (3) in a) mit Bernoulli-Ketten

Wenn man sich aber auf den Standpunkt stellt, dass die Voraussetzungen ‚gegeben' und nicht zu hinterfragen sind, dann zerfällt die von Davies vorgetragene Kritik an der Lösung des Ministeriums gegen Teil (3) der Aufgabe a):

> Wenn sich die Bedingungen IMMER gleich darstellen, dann ist es völlig egal, **wann** man mit dem Datensammeln beginnt.

Dies ist eine fundamentale Eigenschaft von Bernoulli-Ketten im Besonderen und von (echten) Stichproben im Allgemeinen. Man könnte auch – durch Zufall bestimmt – einige Daten eliminieren. Statistiker sprechen von iid-Bedingungen (*independent* = unabhängige, *identically* = identisch *distributed* = verteilte Situationen).

Wieder: Im Sport ist das wahrlich nicht der Fall, aber wir wollen vorderhand das nicht noch einmal hinterfragen.

Wenn es egal ist, dann gehen wir einfach hin zu einem Spiel und warten auf die nächsten Freiwürfe von Nowitzki. Wir wollen sehen, ob er

- nicht mehr als viermal hintereinander trifft – Ereignis $\overline{A}$,
- oder ob er mehr als viermal hintereinander trifft – Ereignis $A$

Klar: $P(A) = p^5 \cdot 1$ und $P(\overline{A}) = 1 - p^5$.

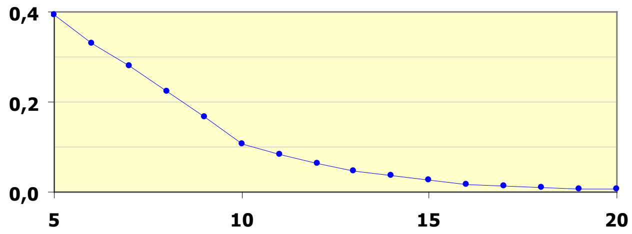
Man braucht den fiktiven Gegner und die künstliche Spielsituation, wie Davies sie schildert, gar nicht. Genauso wenig benötigt man die Angabe der Zahl der Beobachtungen. Man nützt lediglich eine fundamentale Eigenschaft von Bernoulli-Ketten.

### Wenn man die Zahl der beobachteten Freiwürfe vorgibt, verändert sich die Lösung!

Wenn man die Zahl $n$ der Versuche *vorgibt*, so ändert sich die Wahrscheinlichkeit. Beobachtet man nur bis höchstens vier Versuche, so gilt: $P(\overline{A}) = 1$. Je länger man beobachtet, desto eher erhält der Spieler die Chance, doch mehr als viermal hintereinander zu treffen; das bedeutet:

$$P(\overline{A} \mid n) \to 0, \quad n \to \infty.$$

Einfluss der Vorgabe der Dauer der Beobachtung auf die Wahrscheinlichkeit, nicht mehr als viermal hintereinander zu treffen - Simulationsstudie




Wenn man einfach hinkommt und beobachtet, ob Nowitzki es ‚schafft', mehr als viermal hintereinander zu treffen, so ist alles über die 5. Beobachtung hinaus überflüssig; die Lösung stimmt daher mit der Vorgabe $n = 5$ überein.

### Zur Verwirrung von festen, bekannten Modellwerten und ungenauen Schätzwerten

Auch die Kritik von Aufgabe b) von Davies kann man erschüttern. Die Wahrscheinlichkeit, einen Freiwurf zu realisieren, gültig für die ganze Saison,

- kann man natürlich **nach** der Saison als bekannt ansehen
- und nicht, wie Davies es tut, als Schätzwert – einer fiktiven längeren Serie – einstufen.

Die Frage in b) kann man dann so umformulieren:

Vorausgesetzt man kennt diese Wahrscheinlichkeit, ist es dann möglich, die Auswärtsspiele als

- unabhängige Wiederholung desselben Bernoulli-Experiments
- mit eben dieser Wahrscheinlichkeit (für alle Spiele bzw. noch besser für Heimspiele)

zu modellieren? Weichen die beobachteten Trefferzahlen (oder Trefferquoten, das ist gleichwertig) signifikant davon ab oder nicht? – das ist die Frage.

Dabei taucht die Erfolgswahrscheinlichkeit für die Bezugsverteilung (am besten die Heimspiele) als feste Größe auf, für die man nach der Saison einen bekannten Wert hat. Die beobachtete Erfolgsrate oder Erfolgsanzahl für die Auswärtsspiele ist eine empirische Größe, deren beobachteter Wert mit dieser Bezugsverteilung verglichen wird.

## Zentral gestellte Aufgaben

### Folgerungen aus der Kritik an der Kritik

- Die Kritik hält einer Kritik fast weniger Stand als die kritisierte Aufgabe selbst.

Viel schlimmer ist aber die Konsequenz auf die Zukunft von zentral gestellten Aufgaben:

- Die Modellierung muss ‚wasserdicht' sein und darf/kann vom Abiturienten keinesfalls hinterfragt werden (eine zentrale Forderung auch von Davies). Sie muss überdies von allen gleichermaßen klar verstanden werden können.
- Die Textgestaltung ufert zu einer juristischen Aufgabe aus.
- Mehr als Basiskompetenzen wird man so nicht abfragen können.

Mögliche Konsequenzen auf den Unterricht:

- Hinterfragen, warum und ob Ergebnisse relevant sein könnten, hat kaum mehr Platz.
- Übrig bleibt schematisches ‚Anwenden' statistischer Methoden, die mit Basiskompetenzen gelöst werden.

### Teilzentral anstatt zentral organisiertes Abitur

In Österreich wurde erst jüngst für die Zukunft eine im schriftlichen Teil vollzentrale Matura (Abitur) beschlossen. Die Didaktikkommission der Österreichischen Mathematischen Gesellschaft hat sich, unter vielen anderen Vereinigungen, mit einer Resolution dagegen an die Öffentlichkeit gewendet – ohne sichtlichen Erfolg (ÖMG 2009). Kritische Stimmen sind auch aus der Schweiz zu vernehmen (VSG 2009). Sollte eine Matura in Österreich keine ‚lokalen' Anteile enthalten, so ist absehbar, dass sich der Unterricht auf die Vermittlung von Basiskompetenzen konzentrieren wird. Unser Vorschlag daher:

- Zentral gestellte Aufgaben: zur Überprüfung gemeinsamer Sprache und Basiskompetenzen.
- Lokal gestellte Aufgaben: zum sinnvollen Anwenden und Hinterfragen von Ergebnissen.

Für die Überprüfung von Basiskompetenzen und einer gemeinsamen Sprache ist es viel einfacher, klare, einfache und gegen Missverständnisse halbwegs abgesicherte Aufgaben zentral zu erstellen. Dafür sind sie wohl auch gut und geeignet.

Falls sich im lokalen Teil doch Fehler eingeschlichen haben, lassen sich diese viel eher ausmerzen, etwa durch Rückfragen beim Klassenlehrer während der Prüfungsarbeit. Bei zentral gestellten Aufgaben ist ‚online' ein Mut zur Klarstellung von Seiten der Lehrerschaft nicht zu erwarten – wer übernimmt schon das Risiko, eine zentral gestellte Aufgabe möglicherweise falsch zu interpretieren.

## Der Modellierungsaspekt

Weder die Aufgabenstellung an sich noch die folgende Diskussion haben sich daran gestoßen, dass die gewählte Modellierung die anstehenden Fragen – auch wenn die ‚fehlenden' Angaben ergänzt werden – kaum sinnvoll beantworten lässt. Natürlich wollen wir die Lernenden nicht überfordern, wie in der von Davies geäußerten Kritik festgehalten wird:

> „Wenn die Modellierung feststeht, wird die Aufgabe innerhalb des Modells weiter bearbeitet. D. h., es wird angenommen, dass das Modell stimmt. Es ist nicht die Aufgabe des Abiturienten, die Angemessenheit der Modellierung zu hinterfragen."

Unter Modellkritik fallen auch Fragen wie: Unter welchen Annahmen lässt sich die Zahl von Nowitzkis Treffern denn eigentlich als Bernoulli-Kette modellieren? Schnell sieht man ein, dass zwischen Modell und Realität eine Kluft herrscht. Hier hätte das Ministerium die Prüflinge explizit dazu auffordern sollen, die Modellannahmen kritisch zu würdigen und zu kommentieren, warum es damit Probleme gibt. Dies ist nicht erfolgt, was das eigentliche Versäumnis der Aufgabenstellung ausmacht.

Vermutlich sind aber gerade solche Fragestellungen in einer zentral gestellten Prüfung besonders heikel. Wenn diese aber im Abitur nicht gestellt werden (können), ist abzusehen, dass sie auch aus dem Unterricht ‚verschwinden' werden.

Fragen der kritischen Einschätzung der verwendeten Modelle im Unterricht auszuklammern bedeutet aber, Anwendungen auf „Anwendungen" zu reduzieren und sich wesentlicher Chancen eines Unterrichts in Stochastik zu begeben. Gerade die interessanten Modellierungsaspekte, einschließlich des Hinterfragens, warum und ob die Ergebnisse denn auch relevant sein könnten, kennzeichnen einen guten Unterricht in Stochastik.

## Literatur


Bildungsklick (o. J.): *Ministerium bei der Erstellung von Mathe-Aufgaben Zentralabitur überfordert?* http://bildungsklick.de/a/61216/ministerium-bei-der-erstellung-von-mathe-aufgaben-im-zentralabitur-ueberfordert/ (Einsicht: 15. 9. 2008).

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

Davies, P. L. (o. J.): *Stellungnahmen zu einzelnen Aufgaben des Zentralabiturs*. http://www.stat-math.uni-essen.de/lda/zentralabitur.php (Einsicht: 15. 9. 2009).

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

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*. http://www.oemg.ac.at/DK/index.html (Einsicht: 15.9.2009).

VSG (2009): *Die Zukunft des Gymnasiums – Positionspapier des VSG. Verein Schweizerischer Gymnasiallehrerinnen und Gymnasiallehrer*. http://www.vsg-sspes.ch/fileadmin/files/pdf/09.03_Zukunft_Gymnasium_details_d.pdf (Einsicht: 15.9.2009).

## Anschrift des Verfassers

Manfred Borovcnik,
Institut für Statistik,
Alpen-Adria-Universität Klagenfurt
`manfred.borovcnik@uni-klu.ac.at`