
Forschungsprozess und probabilistische Modellbildung – Stochastische Denkweisen

Prof. Dr. Manfred Borovcnik

Zusammenfassung

Die Begriffe der Wahrscheinlichkeitsrechnung sind abstrakt und einer direkten Deutung kaum zugänglich. Die Methoden der beurteilenden Statistik bauen darauf auf und verwenden zusätzlich eine eigene Logik. Häufig werden die Methoden mechanisch angewandt: bei Wahrscheinlichkeit regiert eine primitive Deutung als relative Häufigkeit, die beurteilenden Methoden werden rezeptartig eingeführt. Für echte Anwendungen ist es aber unerlässlich, den Modellbildungsgedanken einzubinden. Das kann auch die Unterweisung verbessern. Wir illustrieren mit Fallbeispielen aus der empirischen „Forschung“ das Potential, über Modellbildung stochastische Denkweisen zu fördern.

1 Modellbildung und Stochastik

Die Debatten um Risiken jedweder Art machen uns bewusst, welch große Verbreitung Wahrscheinlichkeitsmodelle gefunden haben. Sogar Physiker sind bestrebt, kausale Paradigmen durch den Zufall zu ersetzen (Styer 2000). Unsere Gesellschaft ist evidenzbasiert geworden: logische Schlüsse werden in zunehmend unsicherer Welt durch statistische Schlüsse ersetzt.

Das Paradigma empirischer Forschung setzt sich selbst in technischen Wissenschaften durch; nur ein Beispiel dazu: in der Materialforschung

wird unter experimentellen Bedingungen ausgetestet, welche „Mischungen“ sich für bestimmte Zwecke bewähren. Eine evidenzbasierte Orientierung hat auch die Wissenschaften neu gestaltet. So genannte „Soft Sciences“ versuchen, Regelmäßigkeiten oder gesetzesähnliche Beziehungen durch Daten abzusichern.

Wenn es um die Verallgemeinerung von Schlüssen aus empirischen Daten geht, so dienen Methoden der beurteilenden Statistik als Standard. Diese sind in einen allgemeinen Prozess der Erkenntnisgewinnung eingebettet, der mit einer systemanalytischen Erfassung und Modellierung der Ausgangssituation zusammenhängt.

Auch für die Wahrscheinlichkeitsrechnung bietet der Modellierungsgedanke Möglichkeiten, Konzepte neu zu bewerten: So etwa gehen Aussagen

M. Borovcnik ✉

Institut für Statistik, Alpen-Adria Universität Klagenfurt,
Universitätsstr. 65, 9020, Klagenfurt, Österreich

wie „Die Wahrscheinlichkeit beim Würfeln für einen Sechser IST $1/6$ “ am Wesentlichen vorbei. Wahrscheinlichkeit kommt einer realen Situation erst indirekt über eine *Modellierung* zu. Vielmehr ist diese Feststellung Folge eines Modells – der Gleichwahrscheinlichkeit, das bei nur leichten Verletzungen der Symmetrie brauchbare Ergebnisse liefert.

Im Prozess der Modellierung braucht man viel Mathematik, weil man viele Modelle miteinander zu vergleichen hat (in Voraussetzungen wie in Ergebnissen); darüber hinaus sind Kenntnisse des Kontexts unerlässlich. Der Prozess verläuft zyklisch:

- ein vernünftiges Modell für die Situation erarbeiten;
- innerhalb des Modells Lösungen finden;
- diese auf die Situation übertragen;
- die Lösung im Kontext evaluieren; und
- weitere Fragen, die durch das neue Wissen aufgeworfen werden, thematisieren.

Üblicherweise durchläuft man mehrere Zyklen bis zur „Lösung“. Für die Beurteilung eines Modells muss man zwischen Modellebene und realem Problem hin und her übersetzen. Die Übersetzung wird durch typische Muster erleichtert aber auch eingeschränkt. Neben der Interpretation von Wahrscheinlichkeit als relative Häufigkeit (und darauf basierend Erwartungswert als Richt- oder Prognosewert für Mittelwerte in „Stichproben“) gibt es auch die Zusammenfassung der Voraussetzungen einer Wahrscheinlichkeitsverteilung in einer tragfähigen Idee, welche diese Verteilung verkörpert (siehe Borovcnik und Kapadia 2011).

2 Statistische Signifikanz und statistische Experimente

Wir benutzen eine Fallstudie um die Denkweise und das Verfahren des statistischen Signifikanztests zu entfalten. Dabei ist uns aus vielerlei Hinsicht die enge Bindung an den Kontext wichtig. Die Motivation ist wichtig, ja, aber auch, dass man den Prozess der Modellbildung mit dem Kontext immer rückkoppeln kann und die neuen Begriffe

damit in Griffweite bleiben und leichter interpretierbar sind.

Das Verfahren wird dann am Ende eines Modellbildungsprozesses als Zusammenfassung der Vorgangsweise eingeführt. Damit wird eine allgemein gültige Methode begründet, wie man – unter der Einschränkung, dass die Modellierungsschritte auch wirklich passen – eine Sachfrage beantworten kann.

Die Sachfrage „Sind wir besser als die Psychologen es in einem Gesetz formuliert haben?“ wird dabei auf die Ebene des Zufalls verschoben und so umformuliert „Wie ungewöhnlich, nein, wie unwahrscheinlich ist unsere beobachtete Leistung, wenn wir den Zufall als Vergleichsmaßstab heranziehen?“ Letztere Frage ist mit Methoden der beurteilenden Statistik analysierbar. Aus Platzgründen besprechen wir nur Eckpfeiler der Vorgangsweise – auch um eine gewisse Abgeschlossenheit des Themas im vorliegenden Beitrag zu bewahren. Details kann man etwa in Borovcnik und Kapadia 2012 finden.

2.1 Statistische Vorgangsweisen

In der Statistik gibt es mindestens zwei verschiedene Zugänge: beim einen vergleicht man das Ergebnis der vorliegenden Stichprobe mit Hypothesen über einen Parameter und verwendet dabei statistische Tests und Vertrauensintervalle. Der andere ist, Daten in Form einer (unterstellten) Stichprobe zu erzeugen, um dann zielgerichtet zu untersuchen, ob eine Zielvariable von bestimmten Einflussfaktoren abhängt, wobei man solche Abhängigkeiten etwa deswegen beschreiben möchte, um die dahinter steckenden Prozesse besser zu verstehen.

In unseren Fallstudien beziehen wir die „statistischen Modelleure“ in den Prozess der Modellbildung mehrfach ein: die Hypothesen beziehen sich direkt auf sie, die Daten sind persönlich verankert, die Ergebnisse müssen auf sich selbst bezogen und interpretiert werden. So ein Zugang erhöht die Motivation und führt zu einem interaktiven Prozess, in welchem neue Vermutungen

durch Zwischenergebnisse angeregt werden und die Schlussfolgerungen viele neue Fragen aufwerfen. Das verdeutlicht authentisch den Prozess, der zu empirisch fundiertem neuen Wissen führt.

2.2 Der Kontext und die Sachfrage

Wir benutzen ein psychologisches Experiment zum Kurzzeitgedächtnis, um die Teilnehmer direkt in die Analyse miteinzubeziehen. Hierzu wählten wir 15 Worte; jedes war für eine Sekunde auf der Leinwand sichtbar, dazwischen gab es eine kurze Pause. Notizen waren „ausdrücklich untersagt“. Danach sollten die Teilnehmer die Worte aus dem Gedächtnis heraus notieren. Wir folgten einem Vorschlag von Richardson und Reischman (2011).

Der Kontext wurde durch die folgende Frage eingeführt: Wie gut sind wir und wie gut sind Leute im Allgemeinen im Memorieren von Worten („Dingen“)? Wir könnten unsere Ergebnisse mit denen anderer Gruppen vergleichen. Für den wissenschaftlichen Fortschritt gilt das Verhalten *allgemeiner* Gruppen, wie es in gesetzesähnlichen Aussagen formuliert wird, als Vergleichsmaßstab. Solche anerkannten Aussagen sind aus vergleichbaren Experimenten gewonnen worden.

Psychologen haben Experimente zum Kurzzeitgedächtnis in den 1950ern durchgeführt. Miller (1956) resümiert die Ergebnisse in einem Gesetz der magischen 7: Leute können sich mehr oder weniger 7 plus oder minus 2 Items merken. Wichtig ist dabei, dass die Items völlig ohne Zusammenhang sind; jegliche Querbezüge gelten als anerkannter Co-Faktor und würden demnach die Merkaufgabe wesentlich erleichtern.

2.3 Vergleich unserer Ergebnisse mit Hypothesen oder anderen Gruppen

Die Daten stammen aus einem Workshop mit 46 Lehrern in Irland (Details dazu findet man in Borovcnik und Kapadia 2012).

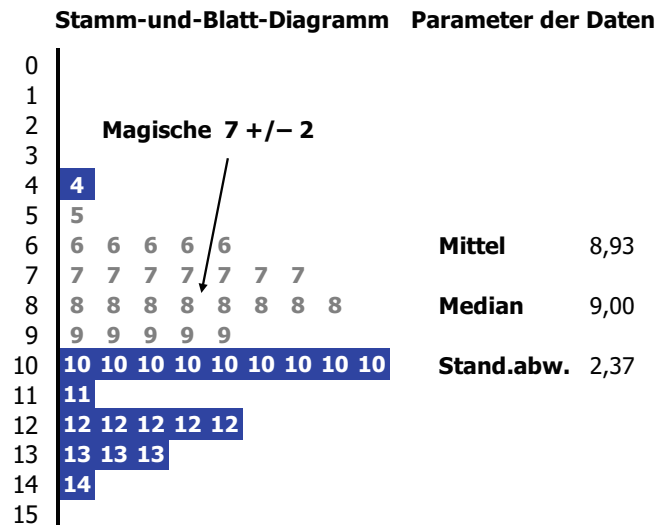


Abb. 1 Stamm-und-Blatt – Ergebnisse, die vom magischen 7 ± 2 abweichen, sind hervorgehoben

Einflussfaktoren für die Zielgröße Merkfähigkeit werden in Abschn. 3 behandelt. Wir starten mit der beurteilenden Statistik; die Einführung des Signifikanztests war das eigentliche Motiv, den Kontext einzuführen. Das Ziel der Fallstudie ist zweierlei: i. Zu beantworten, ob unsere Gruppe besser ist als man nach einem allgemeinen Gesetz zu erwarten hat. ii. Methoden einzuführen, die einen solchen Vergleich ermöglichen.

Unser Zugang ist dadurch gekennzeichnet, dass wir die Verbindung zwischen den formalen Begriffen und dem Kontext aufrecht halten. Das soll die Vorgangsweise besser verständlich machen. Wir verwenden zuerst informelle Wege, um schließlich den Vorzeichentest einzuführen. Wir kürzen aus Platzgründen hier die Entwicklung ab und stellen nur die Eckpunkte dar.

Markieren des normalen Bereichs Die Frage, ob wir besser sind als das Gesetz der magischen 7, kann man schon mit einem Stamm-und-Blatt-Diagramm behandeln (Abb. 1). Wenn wir den Bereich 7 plus minus 2 markieren, so erkennen wir rasch, dass sich etwa die Hälfte in diesem Bereich befindet, während nur ein einziger darunter, die vielen anderen aber darüber sind.

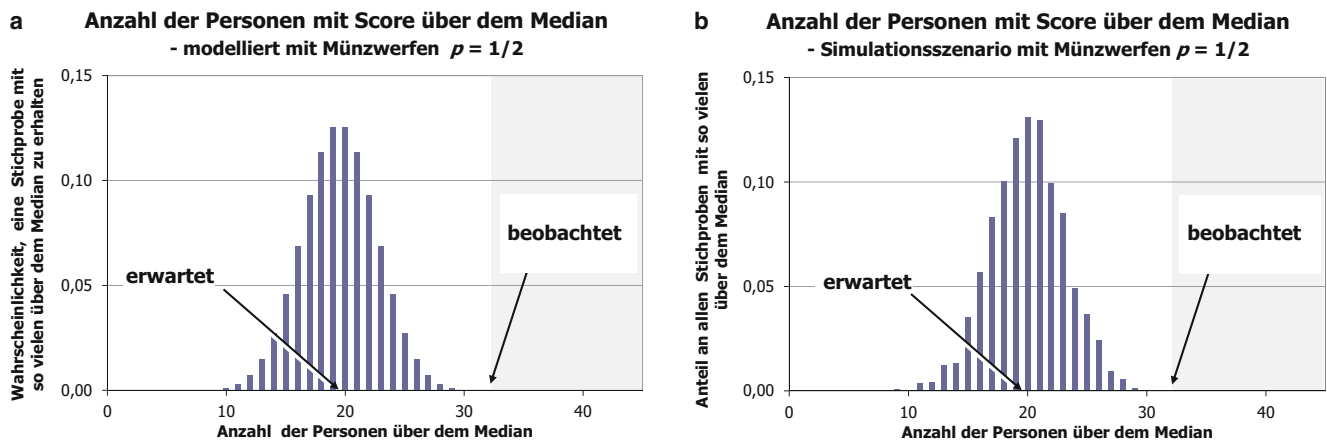


Abb. 2 Potentielle Ergebnisse modelliert durch die Binomialverteilung (a) – Simulierte Ergebnisse (b) – immer unter der Annahme des Münzwerfens, das den „Zufall“ als Vergleich verkörpert

Dies ist ein so deutlicher Hinweis, dass wir als Gruppe besser sind, dass es eigentlich keiner weiteren Verfeinerungen der Methode bedarf.

Modellieren der Bedingungen der magischen 7 – Von Gesetzen zu einer Verteilung „Was kann man unter dem Gesetz der magischen 7 erwarten?“ Mit dieser Frage begeben wir uns ins Reich der Hypothesen. Die Hypothese hinter der magischen 7 ist eine über die Kapazität, sich Worte zu merken. Wir könnten diese Kapazität als *Zufallsvariable* mit einer Normalverteilung modellieren (obwohl sie eigentlich eine diskrete Variable ist) und die Frage dann mit dem so genannten *t*-Test beantworten.

Stattdessen vereinfachen wir die Situation. Wenn wir die 7 als Median unserer zufälligen Kapazitätsvariablen betrachten, so kann eine Testperson darüber oder darunter liegen. Wir lassen gleich alle Personen weg, deren Wert mit 7 übereinstimmt, weil sie ja sowieso nichts zur Klärung unserer Frage beitragen.

Für den Vergleich unserer Gruppe mit dem „Gesetz der 7“ ist nun folgende Annahme grundlegend: Können wir unsere Gruppe als zufällige Stichprobe der theoretischen Kapazitätsvariablen auffassen, wobei der Median dieser Zufallsvariablen 7 beträgt? Wenn das zutrifft, so erfüllen die Daten über (1) bzw. unter dem Median (0) die Anforderungen einer Bernoulli-Kette mit $p = 1/2$,

d. h., wir haben unabhängige Wiederholungen eines dichotomen Experiments. Diese Bedingungen stellt man leicht durch Münzwerfen mit einer fairen Münze dar.

Umsetzen des Modells, um Schranken für „zulässige“ Variation zu bestimmen Um die Frage „Sind wir besser als die magische 7?“ zu beantworten, modellieren wir unsere Daten, *als ob* sie unter den Bedingungen des Münzwerfens entstanden wären. Auf der Basis des Modells der Binomialverteilung mit $p = 1/2$ kann man präzisieren, was man von einer Gruppe, die sich gemäß dem Gesetz der magischen 7 verhält, erwarten kann.

Es zeigt sich, dass es viel stärker überzeugt, die Wahrscheinlichkeit „in Aktion“ zu zeigen, indem man die Bedingungen des Modells *simuliert*. Wir simulieren erst 39 Daten und bestimmen die Anzahl über (und unter) dem Median. Und dann wiederholen wir das Szenario sehr oft (Abb. 2b). Bei 3000 Wiederholungen hat sich kein einziges Mal ein so extrem gutes Ergebnis gezeigt, wie es unsere Gruppe hatte.

Aus der Binomialverteilung berechnet man eine Wahrscheinlichkeit von $4 \cdot 10^{-5}$ (Abb. 2a). Man kann mit *einer* Gruppe mit einem (mindestens) so guten Ergebnis rechnen, wenn man 30.000 vergleichbare Gruppen in einem Experiment testet.

Wir bekommen den Eindruck, dass die Bedingungen des Modells (des Münzwürfens) verletzt sind. Natürlich ist ein so ungewöhnliches Ergebnis (logisch) nicht auszuschließen, aber es legt nahe, dass es besser ist, zu *entscheiden*, dass unsere Gruppe *signifikant* mehr als 7 Worte korrekt abruft.

Die Methode ist als Vorzeichentest bekannt, weil jeder Wert – je nachdem, ob er größer oder kleiner als der Median ist – ein positives oder negatives Vorzeichen bekommt. Werte, die mit dem Median übereinstimmen, heißen Bindungen und werden oft einfach weggelassen; das haben wir ja auch getan.

2.4 Epilog

Ein kleiner Schwenk fehlt noch zur Verallgemeinerung der Vorgangsweise und man hat die Logik des Signifikanztests am Tablett (siehe Borovcnik und Kapadia 2012).

Es ist ganz wesentlich, das gefundene signifikante Ergebnis im Kontext abzusichern. Wenn Ergebnisse signifikant von Hypothesen (allgemeinen Vergleichsgruppen) abweichen, wird das häufig dahingehend missverstanden, dass „etwas“ bewiesen ist. Nein, ein signifikantes Ergebnis ist nur Anlass, darüber nachzudenken, warum unsere Gruppe besser ist. Das erfordert Kenntnisse aus dem Kontext. Deswegen ist uns die Bindung der Einführung der allgemeinen Methode an den Kontext auch so wichtig.

Die Lehrer sind nicht nur Multiplikatoren, die neues Wissen auf ihren Schulen verbreiten. Sie waren auch noch inmitten der Sommerferien auf einem Seminar. Es scheint so, also ob sie tatsächlich gut trainiert sind, neuen Stoff zu memorieren. Eine Übertragung auf alle irischen Lehrer ist durch diese Überlegung aber ausgeschlossen, wengleich es auf einen Versuch ankäme.

Signifikanztests stellen in diesem Sinn einen Filter dar, um im Prozess der empirischen Erkenntnisgewinnung solche Hypothesen herauszufiltern, über die es sich lohnt, vom Kontext her mehr nachzudenken, um Erklärungen dafür zu finden.

3 Statistisches Denken und empirischer Forschungsprozess

In diesem Kapitel erörtern wir, wie man stochastisches Denken auffassen kann. Wir bringen statistisches Denken auch mit dem Prozess der empirischen Forschung zusammen: Wie muss man denken, damit man aus empirischen Untersuchungen Erkenntnisse ziehen kann, die man als evidenzbasiertes Wissen nach innen (innerhalb der Disziplin) und nach außen (in die Gesellschaft) vertreten kann. Wir beschreiben die wesentlichen Merkmale einer Systemanalyse, welche die Aussagekraft der zu erwartenden Ergebnisse sichern soll.

Dazu analysieren wir das Experiment zum Kurzzeitgedächtnis unter systemanalytischen Gesichtspunkten und werten die Daten entsprechend aus. Es wird sich zeigen, dass sich insbesondere die Zeit, wann ein Item präsentiert wird, auf den Erfolg (ob die Leute diese Items korrekt aus dem Gedächtnis abrufen können) auswirkt. Daneben gibt es aber unterschiedliche Strategien, wie Männer und Frauen sich im Experiment orientieren. Der Status der neuen Erkenntnisse ist am besten mit gesetzesähnlichen Regelmäßigkeiten zu umschreiben.

3.1 Stochastisches Denken

Das ist ein altes Schlagwort, welches immer wieder in der Begründung auftaucht, warum man denn Wahrscheinlichkeitsrechnung und Statistik unterrichten soll. Wir wollen über einen vernünftigen Umgang mit einschlägigen Modellen hinausgehen. Stochastisches Denken ist also mehr als einfach Anwenden stochastischer Modelle auf reale (oder artifizielle) Situationen.

Es hat auch etwas mit einem Denken zu tun, das über die Mathematik hinausweist. Es soll hier auch die Ebene der intuitiven Vorstellungen abdecken. Es geht uns hierbei um einen Kurzschluss (im positiven Sinn) zwischen Begriffen und Modellen auf der einen und der Situation auf der anderen Seite. Kurz geschlossen werden die vielfältigen mathe-

matischen Beziehungen, die man vielleicht nicht alle vollständig erfasst hat, von denen man aber eine „Ahnung“ hat. So mag sich Fischbein (1975) die sekundären Intuitionen vorgestellt haben, die sich aus der mathematischen Durchdringung der primären Intuitionen ergeben.

Wir werden gemäß der Trias der Stochastik mit beschreibender Statistik – Wahrscheinlichkeitsrechnung – beurteilende Statistik drei verschiedene Formen ansprechen: Statistische Literalität (statistical literacy), probabilistisches und statistisches Denken.

- Statistische Literalität mag dabei (restriktiv) der beschreibenden Statistik zugeordnet sein und so etwas wie die Fähigkeit bedeuten, Daten auf einem höheren Niveau zu lesen (über Kennziffern und Diagramme).
- Probabilistisches Denken (siehe Abschn. 5) wird mit einem vernünftigen Umgang mit Wahrscheinlichkeit und einschlägigen Modellen in Zusammenhang gebracht. Als Form des *Denkens* sollte es aber über die „rechte“ Handhabung von Modellen hinausweisen: es hat etwas damit zu tun, dass wir eine intuitiv begründete Gewandtheit mit den mathematischen Begriffen erreichen, welche nicht ganz oder nicht immer ausschließlich auf die vollständige Kenntnis der mathematischen Ebene gerichtet ist.
- Statistisches Denken kann analog verstanden sein. Dazu gehört, wie man absichert, wann ein beobachtetes Ergebnis als signifikant einzustufen ist und was signifikant bedeutet. Davon war in Abschn. 2 bei der Entwicklung der Logik des Signifikanztests die Rede.

Wir wollen diese Auffassung von statistischem Denken erweitern; dazu greifen wir eine Idee von Wild und Pfannkuch (1997) auf, welche die statistische Inferenz aus der engen Bindung an Signifikanz lösen und in den Prozess der empirischen Erkenntnisgewinnung einbetten lässt. Wie muss ein Experiment angelegt sein, damit man die Chancen erhöht, daraus Ergebnisse zu bekommen, die einerseits vom Kontext her interessant genug sind und die als neue Erkenntnisse gelten können – die also im Sinne beurteilender Statistik signifikant, d. h., „statistisch abgesichert“ sind.

3.2 Systemanalyse als Angelpunkt eines empirischen Projekts

Der zweite statistische Ansatz ist mit der Suche nach potentiellen Einflussfaktoren verbunden. Dies führt uns in die Systemanalyse; das ist eine systematische Aufbereitung und Modellierung des Kontexts und der Fragestellung.

Systemanalyse oder Modellbildung Was ist die Zielvariable? Welche Faktoren könnten die Werte der Zielvariablen beeinflussen? Welches Skalenniveau haben diese Merkmale? Können wir sie zuverlässig messen? Wie können wir die Ergebnisse unseres Projekts verallgemeinern?

Eine systemische Erfassung der potentiellen Zusammenhänge muss *vor* dem eigentlichen Experiment, das uns die Daten liefert, erfolgen. Die Daten müssen unter dem Gesichtspunkt der anstehenden „Forschungsfrage“ erhoben werden. Wie messen wir die Zielvariable? Ist die Zahl der wiedergegebenen Items ein zuverlässiges Maß der Merkfähigkeit? Wann hat die „Messung“ zu erfolgen, damit sie aussagekräftig ist?

Welche Faktoren beeinflussen die Zielvariable (die Merkfähigkeit in unserem Beispiel)? Über diese Merkmale müssen wir dann auch Daten sammeln, sonst können wir den Einfluss nicht untersuchen. Neben den direkten Einflussfaktoren gibt es so genannte Co-Faktoren (Drittvariable), welche den Einfluss verändern. Solche Merkmale lassen den einen Faktor auf Untergruppen (die durch Werte der Drittvariablen entstehen) anders wirken.

Wenn man Daten über potentielle Störgrößen hat, so kann man deren Einfluss später sogar herausrechnen. Wenn man aber einen Co-Faktor übersehen hat, so kann man – sollte in der Analyse der Daten eine Ahnung auftauchen, dass gewisse Ergebnisse zufolge eines Co-Faktors anders ausfallen als erwartet – die Ergebnisse nicht mehr korrigieren. Sie können daher kaum verallgemeinert werden (weil der Verdacht ja bleibt) oder müssen in späteren Studien allenfalls berichtigt werden. Solche Drittvariable (über die man keinerlei Daten verfügt) heißen im Jargon auch Confounder.

Im Gedächtnistest etwa könnten wir trotz aller Signifikanz das Ergebnis vergessen, wenn etwa folgende Confounder wirksam gewesen sind: Die Worte hängen zusammen, was die Aufgabe des Merkens erheblich erleichtert. (Einige) Testpersonen wussten die Worte schon vorher. Die Testpersonen haben voneinander abgeschaut. Die Interpretation von Exzellenz wäre dann fehl, weil die eigentliche Leistung durch die Confounder mitverursacht wäre. Es gilt, Störgrößen systematisch zu erfassen und entweder Daten dazu aufzuzeichnen oder das Experiment so zu kontrollieren, dass gesichert ist, dass sie ausbleiben.

Die Faktoren im Gedächtnisexperiment Zielgröße ist die Leistung im Merkfähigkeitstest. Wir wollen sie mit dem Score der richtig wiedergegebenen Worte unmittelbar nach dem Vorführen der Worte messen.

Wir benennen summarisch einige wichtige Einflussgrößen. Normalerweise ist dieser Prozess in der Praxis besonders vage und muss dazu noch unter Einschränkungen von Zeit und Geld ablaufen:

Die Personen Alter, Geschlecht, Bildung, Erfahrung mit Mnemotechniken, Vertrautheit mit Memorieren.

Items Einzelne Items haben einen Kontext (sind lustig, mögen „grausliche“ Assoziationen hervorrufen etc.), sind lang oder kurz, mögen fremd wirken (Fremdworte); auch wenn es so nicht beabsichtigt ist, mögen die Worte zusammenhängen. Die zeitliche Reihenfolge mag wichtig sein.

Testsituation Einige mögen die Situation ernst nehmen, andere sind wenig interessiert, was beliebige Antworten hervorrufen kann. Die Tageszeit, das Medium, wie man die Worte präsentiert (auditiv oder visuell) können eine Rolle spielen.

Wechselweise Abhängigkeiten Zeit und Kontext können einander beeinflussen oder auf verschiedene Personen (weiblich, männlich) unterschiedlich wirken.

3.3 Analyse der Wirkung einzelner Faktoren auf die Zielgröße Erfolg

Wir analysieren den Einfluss des Kontexts, der Zeit und des Geschlechts auf den Erfolg im Merkfähigkeitsexperiment. Weil der Zeitpunkt der Präsentation keinen linearen Einfluss auf die Erfolgsraten hat, führen wir als weitere Variable die Tiefe der Präsentation ein. Ziel ist es, die gefundenen Einflüsse durch einfache Beziehungen zu beschreiben.

Beeinflusst der Kontext der Worte den Erfolg beim Abrufen aus dem Gedächtnis? Für eine qualitative Variable gibt eine Rangliste einen ersten Überblick (Abb. 3). Danach hat Kontext der Worte einen Einfluss, denn die Scores schwanken beträchtlich.

Die Worte *rigging*, *ear* und *octopus* führen die Liste an; *seed* und besonders *legend* sind weit abgeschlagen. *Octopus* könnte man noch oben erwarten; für die anderen Worte ist ihr Einfluss weniger klar. Der mittlere Teil erstreckt sich über einen Bereich von 36 bis 65 % ohne sichtbare Gruppierung.

Ja, Kontext spielt eine Rolle, aber das Muster ist zu vielfältig, um daraus weitergehende Schlüsse zu ziehen. Das mag gleichzeitig ein Hinweis sein, dass die Worte doch ziemlich neutral und unabhängig gewählt worden sind, was den Test ja eigentlich zuverlässiger machen würde.

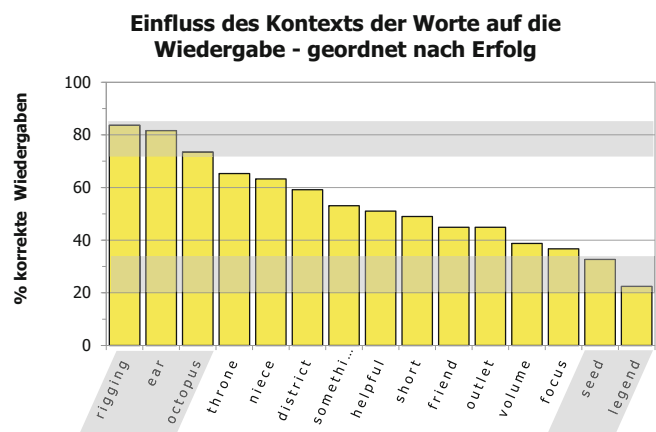


Abb. 3 Rangliste der Worte nach Erfolg – extreme Werte und Spielraum hervorgehoben

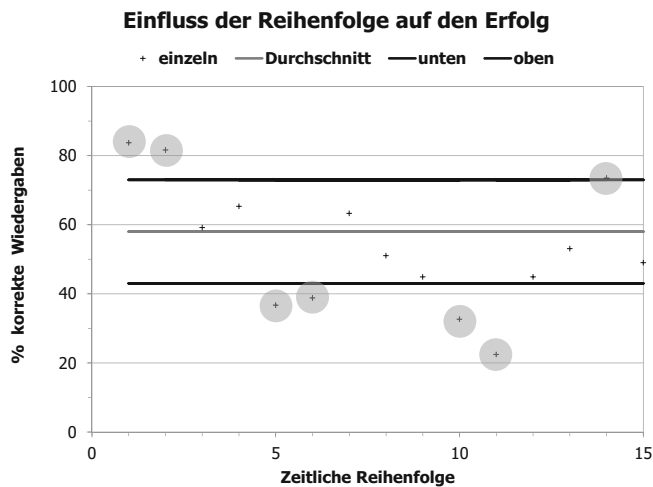


Abb. 4 Erfolgsrate abhängig von der zeitlichen Präsentation – extreme Werte markiert

Zeitpunkt der Präsentation beeinflusst den Erfolg beim Abrufen aus dem Gedächtnis Es entspricht dem Hausverstand, dass der Zeitpunkt, wann die Items präsentiert wurden, den Erfolg beim Abrufen aus dem Gedächtnis beeinflusst. Man konnte im Seminar mitverfolgen, wie sich die Teilnehmer von Anfang an ins Zeug legten und die bereits gesehenen Worte memorierten. Der letzte Eindruck zählt, wie man so sagt; daher sollten auch die letzten Worte bessere Chancen haben. Das macht es plausibel, den Einfluss von Zeit (der zeitlichen Anordnung der Items), aber auch der Tiefe eines Worts (minimaler Abstand von Anfang und Ende der Präsentation) auf den Erfolg der Wiedergabe zu untersuchen.

Die Darstellung der Erfolgsrate in Abhängigkeit von der Zeit (Abb. 4) zeigt erwartungsgemäß ein deutliches Muster; dabei sind mittlere Zeitpunkte mit kleinen Erfolgsraten verbunden. Die Punktwolke zeigt auch, dass es sinnvoll wäre, den Erfolg gegen die Variable Tiefe darzustellen.

Die Ergebnisse sind aber nicht nur im Einklang mit dem Hausverstand, sie können auch durch psychologische Beziehungen erklärt werden. Die hohe Anfangskonzentration wird zur Mitte des Experiments hin abfallen, während sich die letzten Worte noch im Gedächtnis abheben, wenn die Teilnehmer die Worte niederschreiben.

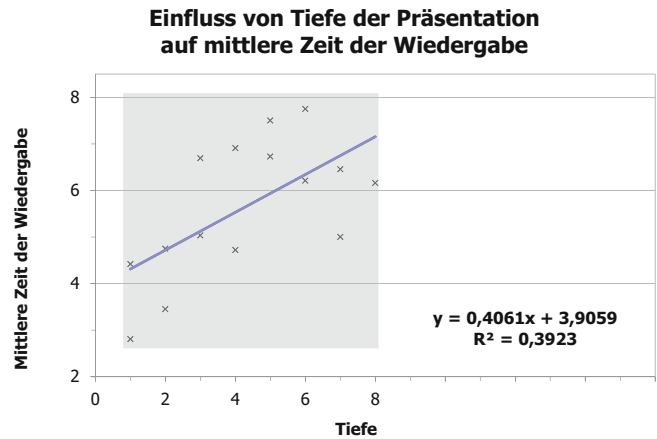


Abb. 5 Einfluss der Tiefe auf die Zeit des Abrufens – standardisierte Punktwolke

Hat der Zeitpunkt der Präsentation einen Einfluss auf die Reihenfolge des Abrufens? Memorieren die Teilnehmer die Worte in der Reihenfolge, in der sie präsentiert werden und – als Folge davon – rufen sie diese in dieser Reihenfolge aus dem Gedächtnis ab? Die zeitliche Reihenfolge des Abrufens wurde aus dem Arbeitsblatt rekonstruiert.

Wir nehmen als unabhängiges Merkmal die Tiefe der Darbietung und als abhängiges den Zeitpunkt der Wiedergabe. In der Punktwolke sieht man ein deutliches Muster der Gleichsinnigkeit; dies wird durch eine Korrelation von 0,63 bestätigt (Abb. 5).

Um die Korrelation aus Punktwolken ablesen zu können, ist es wesentlich, die Achsen zu skalieren, sodass die Punkte in etwa innerhalb eines Quadrats zu liegen kommen. (Tatsächlich kann man den graphischen Eindruck durch Skalieren der Achsen beliebig manipulieren.)

Weitere Einflussfaktoren Auf den Arbeitsblättern war eine beliebige dreigliedrige Kennzahl einzutragen. Kann man deren Eigenschaften mit dem Erfolg in Verbindung bringen? Die Korrelation der ersten Ziffer mit dem Score ist mäßig; auch die Punktwolke zeigt kein Muster. Interessanter ist, dass die Eigenschaft Primzahl einen kleinen, aber nicht-signifikanten Einfluss hat. Das erspart uns, nachzudenken, wie und wieso eine Präferenz

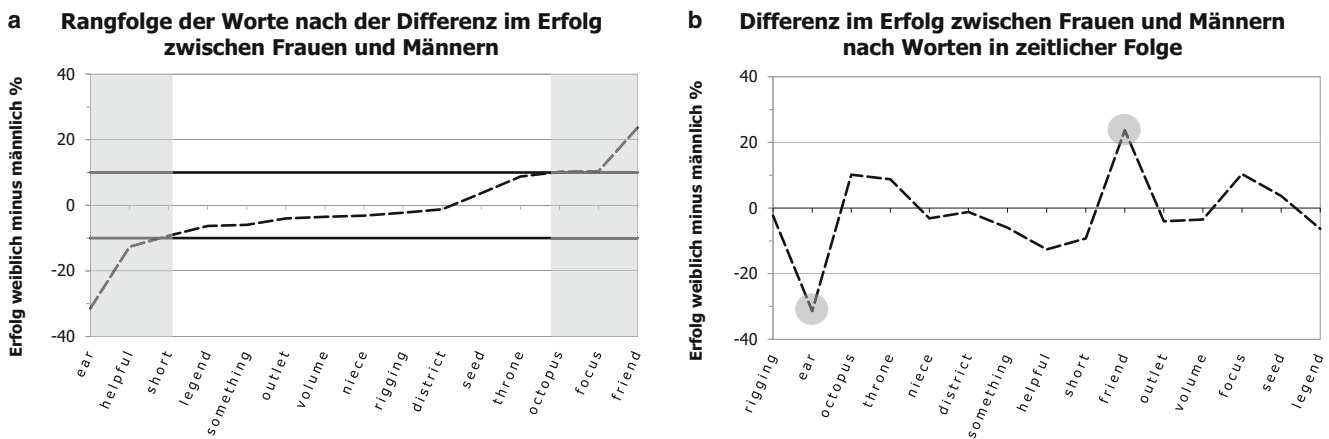


Abb. 6 Geschlechtsunterschiede – nach Größe der Differenz in den Erfolgsraten (a) und nach Worten in zeitlicher Reihenfolge (b)

für Primzahlen mit der Erfahrung oder der Fähigkeit, sich Worte zu merken, zusammenhängt.

3.4 Interaktionen zwischen Faktoren

Geschlecht verändert, wie andere Faktoren den Erfolg beim Abrufen beeinflussen Zunächst untersuchen wir den Unterschied im Erfolg nach Geschlecht. Mit den verfügbaren Daten (25 Frauen; 18 Männer; von 3 Personen konnte die Angabe zum Geschlecht nicht rekonstruiert werden) haben wir eine mittlere Anzahl von korrekt wieder gegebenen Worten von 9,25 und 8,06; der Unterschied von 1,19 Items zugunsten der Frauen ist aber nicht signifikant ($t = 1,54$).

Berücksichtigt man die fehlenden Daten (6 weibliche und 1 männlicher Teilnehmer hatten in diesem Teil das Arbeitsblatt leer) und untersucht einen worst case, so bricht der Unterschied auf 0,36 Items zusammen. Damit haben wir nicht nur ein Ergebnis, das nach den Regeln der Kunst empirischer Forschung nicht abgesichert ist (ist nicht signifikant), es scheint auch sachlich kaum relevant zu sein; der Unterschied entspricht knapp 4 % der durchschnittlichen Leistung.

Es gibt mehr zu Geschlechtsunterschieden zu sagen. Wirkt sich der Kontext unterschiedlich auf Männer und Frauen aus? Welche Worte hatten den größten Unterschied im Erfolg der beiden Grup-

pen? Wirkt sich die Zeit der Platzierung der Worte unterschiedlich aus? Kann dieser Einfluss durch den Kontext der Worte verändert (verschleiert) werden?

Das Ranking der Worte nach der Differenz im Erfolg (weiblich minus männlich; Abb. 6a) zeigt, wo die Geschlechter *im Vergleich zueinander* besonders gut oder schlecht abgeschnitten haben: größter Vorsprung bei Männern bei den Worten (*ear*, *helpful*, *short*); größter Vorsprung der Frauen bei (*friend*, *focus*, *octopus*).

Ob diese Differenz durch die Bedeutung der Worte (Kontext) oder durch die zeitliche Reihenfolge bedingt ist, kann man anhand der Abb. 6b untersuchen; diese deutet an, dass Frauen am Beginn weniger Erfolg hatten (Tiefe 2, *ear*) und in der Mitte erfolgreicher waren (Tiefe 7, *friend*).

Ergebnis der Untersuchungen der wechselweisen Abhängigkeiten als Hypothese Männer und Frauen verfolgen beim Memorieren von Worten verschiedene Strategien, welche sich unterschiedlich an Zeit und Kontext orientieren: Männer verhalten sich im Experiment eher sequentiell, Frauen sind stärker am Kontext der Worte ausgerichtet.

Weitere, zielgerichtete Untersuchungen hierzu könnten das Verhalten der Testpersonen – äußerlich – aufzeichnen oder durch strategisches Variieren von Kontext und Position der im Experiment verwendeten Worte zu erforschen suchen.

3.5 Bewertung der neuen Erkenntnisse

Die gewonnenen Ergebnisse haben den Status von gesetzesähnlichen Beziehungen. Ehrenberg (1981) sieht es als vornehmliches Ziel empirischer Forschung, solche „Gesetze“ aus empirischen Untersuchungen herauszufiltern und zu untersuchen, unter welchen zusätzlichen Bedingungen sie „verallgemeinert“ werden können.

Allerdings müssen wir die Gültigkeit sowieso auch schon wieder einschränken, denn wir haben ja mehrfache Tests angewendet; dabei ist es normal, dass dann einige Besonderheiten zu erkennen sind. Wir müssten – zur Korrektur – einen strengeren Maßstab anwenden, ab wann wir ein Ergebnis als signifikant anerkennen. Wir haben aber noch einen zweiten methodisch angreifbaren Umstand: wir haben uns durch Zwischenergebnisse zu weiteren Analysen anregen lassen. Eigentlich sollten wir schon vor der Erhebung einen Plan haben, was alles zu untersuchen ist.

Viele Ergebnisse empirischer Forschung haben keinen größeren Anspruch auf Verallgemeinerung als unsere. Replikation wäre ein magischer Schlüssel, um die Beziehungen zu erhärten oder als isolierte Phänomene zu erkennen. Allerdings werden Replikationen eher selten durchgeführt.

Der Ansatz, Ergebnisse durch Replikation abzusichern, würde auch mit der Logik der Forschung von Popper (1935) in Einklang stehen. Daher sollten wir besser von Hypothesen sprechen, die sich aus unseren Untersuchungen ergeben. Diese warten darauf, dass sie durch weitere Forschung erhärtet werden. Das ist die eigentliche Aufgabe und Ausrichtung empirischer Forschung, welche evidenzbasiertes Wissen anhäuft.

Die abschließende Diskussion der Ergebnisse der Statistiker mit den Experten aus dem Kontext sollte noch folgende wesentliche Punkte umfassen: Evaluation der Ergebnisse im Kontext, Integration der Beziehungen in ein theoretisches Wissensgebäude, Erörterung von möglichen Widersprüchen der Ergebnisse mit langfristiger Erfahrung sowie eine Synopse von Fragen, welche durch das neue Wissen entstanden sind.

4 Weitere Fallstudien

Wir haben andere Aktivitäten ähnlich zum Gedächtnistest ausgearbeitet und in Seminaren erprobt. Ziel ist es, Eigenheiten von Wahrscheinlichkeitsbegriffen und statistischen Verfahren besonders gut hervorstechen zu lassen. Die folgenden Vorschläge für Projekte weisen auch einen hohen Grad an Interaktivität und Selbstbezug auf, aus dem sich ‚drängende‘ Fragen ergeben, auf denen man Hypothesen aufbauen kann, welche dann durch die weitere Bearbeitung geklärt werden sollen. Wir beschreiben einige der Aktivitäten kurz und verweisen auf Quellen für mehr Details.

Spaghetti brechen ist ein Experiment mit einem verblüffenden Ergebnis, da die relativen Häufigkeiten völlig von einer naiven Modellwahrscheinlichkeit abweichen, was anzeigt, dass etwas falsch gelaufen sein muss. Es führt direkt in grundsätzliche Diskussionen über die Frage „Was ist Zufall?“ und zeigt, dass naives Verhalten ohne jegliche Absicht oder Präferenz gar nichts mit Zufall gemeinsam haben muss.

Personen zu bitten, eine beliebige Codezahl mit drei Ziffern zu notieren, ist Ausgangspunkt einer „Studie“ zu Präferenzen für Zahlen. Die Ergebnisse zeigen, dass starke Präferenzen vorhanden sind, wobei Präferenz und zufällige Wahl in einer Art Gegensätzlichkeit stehen. Präferenz wird dann zuerkannt, wenn der Zufall als Erklärung „zu wenig Aussagekraft hat“. Muster findet man immer, die Frage ist, ab wann man ein Muster als eine Präferenz anerkennen soll; mit anderen Worten, was meinen wir mit einem *signifikanten* empirischen Befund?

Der Placeboeffekt besagt, dass Menschen eine positive Wirkung (Heilung oder Nachlassen von Schmerzen) empfinden, wenn sie nur glauben, dass sie eine (vielversprechende) Therapie bekommen. Ein psychologischer Wirkmechanismus hilft heilen? In diesem Experiment erklären wir das wohlbekannte Phänomen durch ein einfaches probabilistisches Arrangement, welches im Gegensatz zu einem psychologischen Gesetz nur auf den reinen Zufall zurückgreift.

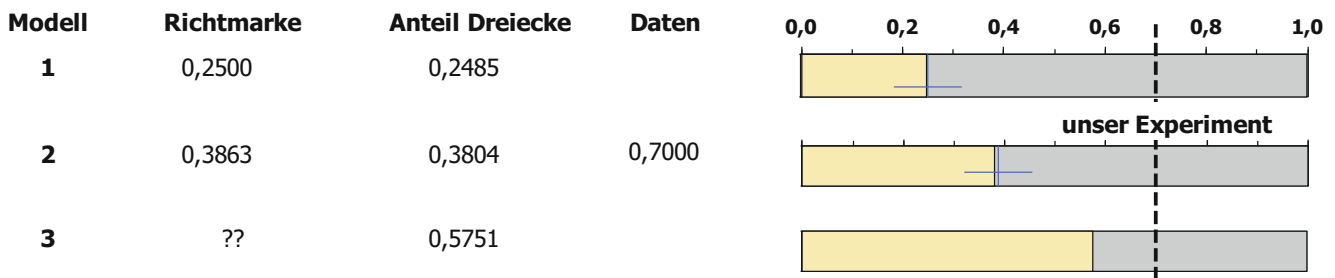


Abb. 7 Simulationsstudie zum Vergleich der Modelle – Streubreite der Simulation durch *Kreuzchen* für den Richtwert sichtbar gemacht; empirische Ergebnisse durch *gestrichelte Linie* markiert

4.1 Spaghetti brechen

Relative Häufigkeiten und Wahrscheinlichkeiten sind wie Zwillinge. Je mehr Daten man hat, umso näher liegen die beiden zusammen. Will man die Schätzung von Wahrscheinlichkeiten verbessern, nimmt man mehr Daten. Der Zusammenhang spiegelt sich im Bernoullischen Gesetz der großen Zahlen. Die stillschweigende Annahme dahinter ist, dass die Daten aus einer Zufallsstichprobe stammen, d. h., dass sie durch unabhängige identische Zufallsexperimente realisiert werden. Im folgenden Experiment liegen diese Werte so weit auseinander, dass man die Situation überdenken und neu modellieren muss. Übrig bleibt: Daten können wertlos sein, wenn sie Voraussetzungen verletzen.

Ein Experiment entwickeln Das folgende Experiment zum Brechen von Spaghetti (Kataoka et al. 2009) soll das Verständnis von Zufall klären. Wir brechen Spaghetti in drei Teile. Danach prüfen wir, ob wir mit den Teilen ein Dreieck bilden können oder nicht. Üblicherweise ergeben sich in den Gruppen mehr als 70 % Erfolgsraten. Die Teilnehmer bekommen je drei Spaghetti, die sie einfach (zufällig?) brechen sollen. Erst dann teilt man ihnen mit, dass sie damit ein Dreieck formen sollen.

Wenn wir die Spaghetti ohne Absicht brechen, sollten wir den Vorgang mit reinem Zufall modellieren können. Wir brauchen dazu zwei Zufallszahlen aus dem Einheitsintervall. Entweder legen die beiden Zahlen direkt die Längen der Bruchstücke fest (Modell 1) oder wir modellieren den Vorgang hierarchisch, erst das erste Teil zufällig brechen, dann einen Teil nehmen und wieder

zufällig brechen (Modell 2). Die Modelle liefern Erfolgswahrscheinlichkeiten von 0,25 bzw. 0,3863 (für die Details siehe Borovcnik und Kapadia 2012).

Dass beide Werte so weit weg von den festgestellten Erfolgsraten sind, wirkt sich regelrecht wie ein Schock aus. Offensichtlich brechen die Teilnehmer die Spaghetti eben nicht zufällig.

Beim Versuch, das Verhalten besser zu modellieren, bemerken wir, dass wir beim zweiten Mal selten das kleinere Stück brechen; ferner ist keines der Teilstücke allzu klein, d. h., wir vermeiden es, zu kleine Anteile abzubrechen. Wir modifizieren Modell 2 und nehmen einen weiteren Parameter q (Mindestanteil beider Bruchstücke) ins Modell auf. Modell 3 legt q mit $1/10$ fest. Abb. 7 zeigt die Ergebnisse einer Simulationsstudie für alle Modelle.

Hintergrund Zufall existiert nicht, hat de Finetti (1974) in seiner Fundierung des subjektivistischen Ansatzes festgestellt. Wir möchten ergänzen: Zufall ist nur eine Sicht auf die Welt. Daraus folgt, dass Wahrscheinlichkeit und abgeleitete Begriffe *Modelleinheiten* sind. Natürlich gibt es eine starke Identifikation mit dem Zufall, wenn Situationen sich durch Fairness, Fehlen kausaler Einflüsse, Fehlen von Mustern etc. auszeichnen. Und wenn solche Situationen wiederholbar sind, bilden relative Häufigkeiten nicht nur eine tragende Interpretation sondern dienen auch zur Schätzung von Wahrscheinlichkeit. Die Schwierigkeit liegt jedoch im Detail: nicht immer sind die genauen Bedingungen wirklich klar, unter denen relative Häufigkeiten nützlich sind.

Auf den ersten Blick ist es für viele Teilnehmer völlig inakzeptabel, dass ihre Ergebnisse so stark vom Zufall abweichen. Es war ihnen gar nicht bewusst, dass sie so ausgeprägte Präferenzen haben. Wenngleich wir nun Spaghetti anders als durch Zufall brechen, ist bemerkenswert, dass wir das Modell durch den weiteren Parameter Mindestanteil doch ein ordentliches Stück verbessern können. Modell 3 mit $q = 1/10$ liegt mit knapp 60 % nicht mehr so weit weg von den Erfolgsraten von 70 %; mit $1/4$ kommen wir noch näher hin, aber das scheint doch weniger plausibel zu sein.

Das Experiment ladet zum Nachdenken ein, dass augenscheinlich zufällige Prozesse gar nichts mit Zufall zu tun haben müssen. Will man relative Häufigkeiten zur Schätzung von Wahrscheinlichkeiten verwenden, so muss der Prozess des Entstehens der Daten gewissen Kriterien genügen (unabhängige, identische Zufallsexperimente). Viele Daten werden verwendet ohne jegliche Rechtfertigung. Die Parabel von „Stichprobe und repräsentativ“ wird wie ein Mantra wiederholt, ohne die Voraussetzungen zu prüfen.

Zufall ist nur ein Baustein zum *Modellieren*. Sogar die Grundlagenphysiker sind gelegentlich sehr schlampig (im Gebrauch der Sprache?), wenn sie uns erzählen, dass sie mit der Urknalltheorie beweisen, dass der Ursprung des Universums zufällig ist.

Wir sollten die Liste, was Zufall ausmachen könnte, modifizieren: wir assoziieren mit Zufall gelegentlich „keinerlei Präferenzen“ oder „keine Absicht“ und erwarten in diesem Fall Gleichwahrscheinlichkeit. Da viele Leute überzeugt sind, dass sie die Spaghetti ohne Absicht oder Präferenzen brechen, glauben sie ganz stark, dass die Ergebnisse zufällig SIND (sogar gleichwahrscheinlich). Umso größer ihre Überraschung, wenn sie feststellen, dass ihre Ergebnisse so stark vom Zufall abweichen. Diesen kognitiven Konflikt kann man zur weiteren Klärung der Konzepte nutzen.

4.2 Präferenzen für Zahlen

In der Fallstudie mit den Spaghetti war die Frage, ob Leute sich wie zufällig verhalten, wenn sie Spaghetti brechen. Hier geht es um die Frage, ob wir beim Auswählen von Zahlen psychologischen Mustern – Präferenzen – folgen oder ob wir unser Verhalten durch reinen Zufall erklären können. Immer interessant ist, ob Drittvariable wie Geschlecht einen Einfluss haben.

Hintergrund Wählen Menschen Zahlen willkürlich oder haben sie Präferenzen? Wir werden willkürlich als „ohne Absicht“ durch reinen Zufall modellieren. Wir könnten Teilnehmer auffordern, Zahlen so zufällig wie möglich zu wählen. Bekannt ist, dass sie längere Runs von gleichlautenden Zahlen sowie deutliche Muster vermeiden. Wir schlagen eine offenere Aktivität vor und fragen simpel nach einer Codezahl aus drei Ziffern.

Einige Ergebnisse Wir beziehen uns auf Daten aus dem Seminar in Limerick (Abb. 8). Die Zahlen der Frauen streuen über die gesamte Spannweite, Männer bevorzugen die Außenbereiche. Für einzelne Ziffern gibt es bei Frauen eine Präferenz für 1 und 3, mit 48 % für 1 3 7 als Gruppe; Männer wählen bevorzugt die 7, wobei 80 % auf die Gruppe 7 3 1 (in dieser Reihenfolge der Häufigkeit) fallen. Für die einzelnen Positionen (erste, zweite, dritte) sind die Ergebnisse ähnlich. Nur die 4 in der zweiten und die 5 in der dritten Position ist bei Frauen stärker vertreten.

Ähnlich wie im Gedächtnistest kann man die Daten auf Signifikanz prüfen. Wir vergleichen zwei gegensätzliche Hypothesen: rein zufällige Auswahl aus dem Bereich von 000 bis 999 (*Nullhypothese*) gegen ausgeprägte Präferenzen. Wir geben dabei den p -Wert an, das ist die Wahrscheinlichkeit, dass man mindestens so „extreme“ Daten beobachtet, wenn die Nullhypothese zutrifft: Primziffern ($p = 10^{-4}$), 1 3 7 ($p = 10^{-6}$), 7 ($p = 0,02$), Primzahlen (als Eigenschaft der Codezahl, $p = 0,02$).

a		b	
Weiblich	Männlich	Weiblich	Alle Männlich
00	0*	0	0
147 132 131	10 121 121	1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1
198	1* 172	2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2
231 211	20 233	3 3 3 3 3 3 3 3 3 3 3 3 3 3	3 3 3 3 3 3 3 3 3 3 3 3
257	2* 291	4 4 4 4 4 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4 4 4
349 347 335 333	30 327 341 347	5 5 5 5 5 5 5 5 5 5 5 5	5 5 5 5 5 5 5 5 5 5
375 372 369 365	3* 371 397	6 6 6 6 6 6 6 6 6 6	6 6 6 6 6 6 6 6 6 6
437 411	40	7 7 7 7 7 7 7 7 7 7 7 7	7 7 7 7 7 7 7 7 7 7 7 7
475 461	4*	8 8 8 8 8 8 8 8 8 8	8 8 8 8 8 8 8 8 8 8
542 518 514	50	9 9 9 9 9 9 9 9 9 9	9 9 9 9 9 9 9 9 9 9
599	5*		
615	60		
685	6* 653		
747 713 702	70 732 743 745		
	7* 757 789		
842	80		
	8* 876		
	90 912		
	9*		

Abb. 8 Präferenzen für die Codezahlen nach Geschlecht getrennt – **a** alle 3 Ziffern; **b** einzelne Ziffern

Caveat Wir haben Tests mehrfach angewendet; die Analyse war auch durch Zwischenresultate angeregt. Das ist methodisch fragwürdig. Viele Ergebnisse empirischer Forschung sind allerdings so zustande gekommen und haben keinen besseren Anspruch auf Verallgemeinerung als unsere Präferenzstudie.

Auch hier wäre Replikation angebracht, um Artefakte von Evidenz zu trennen. Hypothesen, die sich in wiederholten Studien bewähren, gewinnen an Glaubwürdigkeit.

4.3 Placeboeffekt und das Phänomen der „Regression zur Mitte“

Der Placeboeffekt ist ein anerkannter psychologischer Wirkmechanismus. Allein die Erwartungshaltung hat messbare Folgen, auch wenn überhaupt keine Einflussnahme erfolgt. Regression zur Mitte ist ein Phänomen, wonach die Besten in einem „Wettbewerb“ bei Wiederholung zwar noch immer über der Mitte liegen, aber näher zum Zentrum rücken. Beide Phänomene kann man durch Vergleich mit einem Zufallsexperiment als Artefakt bezeichnen.

Hintergrund Allein die Erwartung, eine medizinische Therapie oder Intervention zu erhalten, hat

einen messbaren Heilungseffekt. Empirische Studien zum Nachweis der Wirksamkeit eines neuen Medikaments oder einer Therapie müssen sich als besser als Placebo erweisen. Placebo ist eine Substanz, die äußerlich dem Testmedikament (Verum) gleicht, aber keine medizinische Wirksubstanz enthält. Placebo wird unter exakt denselben Bedingungen verabreicht wie Verum. In medizinischen Studien sind Teilnehmer und behandelnde Ärzte verblindet; niemand weiß, was verabreicht wird. Zusätzlich werden die Teilnehmer zufällig einer der Gruppen Placebo (Kontrolle) oder Verum (Behandlung) zugeordnet.

Das hat sich zum goldenen Standard in der medizinischen Statistik herausgebildet, um zu garantieren, dass die Personen in beiden Gruppen möglichst ähnlich sind und dass ihre Erwartungen keinerlei Einfluss auf die messbare Wirkung haben. Das Design erlaubt es, die unterschiedliche Wirkung allein als Folge der Behandlung zu interpretieren. Die Daten werden ähnlich wie im Gedächtnisexperiment verwendet, um die Frage zu beantworten: „Ist Verum – signifikant – besser als Placebo?“

Regression zur Mitte ist ein weiteres „Gesetz“. Wir geben ein Beispiel: Wenn statistische Einheiten unter zwei Bedingungen beobachtet werden – etwa vor und nach einer Unterrichtseinheit – so werden die besten vorher zum Mittel „zurück-

schreiten“, d. h., ihre Leistung nachher wird weniger weit vom Mittelwert entfernt sein als zuvor. Für diejenigen, die vorher zu den schlechtesten zählten, liegen die Verhältnisse gegengleich.

Historisch ist das Beispiel mit den Vätern und Söhnen berühmt geworden. Damit wollten Galton und Pearson in der zweiten Hälfte des 19. Jahrhunderts in einem groß angelegten Forschungsvorhaben die Vererblichkeit von Intelligenz zeigen. Da die Konstrukte zur Messung von Intelligenz erst später entwickelt wurden, versuchten die Forscher, die Vererblichkeit mit direkt messbaren Merkmalen wie der Körpergröße zu „beweisen“. In Freedman et al. (2007), oder MacKenzie (1981) findet man mehr dazu. Väter, die zu den größten zählten, hatten nun Söhne, die noch immer über dem Mittelwert lagen, aber weniger extrem waren als ihre Väter. Dies wurde als Regression (Rückschreiten) bezeichnet. Die Kennziffer, welche den Grad der Erblichkeit messen sollte, wurde *Regressionskoeffizient* genannt.

Entwicklung eines Experiments Folgendes Experiment wurde von Dubben und Beck-Bornholdt (2010) vorgeschlagen: Wir werfen 42 Würfel ein erstes Mal und betrachten nur die extremen Ergebnisse 6 (Top) und 1 (Flop); nur mit diesen werfen wir ein zweites Mal. Ein einzelnes Doppelexperiment mag als Mittelwert 3,64 für Flops und 3,50 für Tops ergeben. Zufall! Die Flops sind im Durchschnitt sogar besser als die Tops. Wiederholen wir das Doppelexperiment oft, werden die Tops des ersten im Durchschnitt im zweiten schlechter während sich die Flops verbessern.

Damit ist die Regression als Verhalten bei stochastisch *unabhängigen* Experimenten erklärt – keine Rede von einem genetischen Gesetz. Für den Placeboeffekt merken wir an: Patienten gehen gerade dann zum Arzt, wenn es ihnen schlecht geht (Flop beim ersten Wurf) und es geht ihnen dann besser, egal was der Arzt verschreibt – das bessere Ergebnis beim unabhängigen zweiten Wurf.

5 Probabilistisches Modellieren

Die Rolle von Wahrscheinlichkeit innerhalb der Curricula zu stärken (Borovcnik 2011b) ist eine wichtige Forderung, weil man erst damit statistische Methoden verstehen kann. Die Kennziffern von statistischen Tests sind – indirekt – meist mit *bedingten* Wahrscheinlichkeiten verbunden. Mit dem Modellbildungsgedanken kann man Wahrscheinlichkeit sinnvoller gestalten.

Hier kommen in natürlicher Weise auch andere – subjektive – Interpretationen ins Spiel. Damit werden allzu enge Auffassungen von Wahrscheinlichkeit als relative Häufigkeit aufgebrochen und sowohl die Anwendungen als auch das Verständnis verbreitert. Die Bindung an empirische Häufigkeiten ist wohl in einer objektiven wissenschaftlichen Auffassung verankert. Aber, erstens ist jede Anwendung bis zu einem gewissen Grad subjektiv und zweitens wird durch die Bindung auch die Interpretation von Methoden der beurteilenden Statistik fragwürdig.

Wir beschreiben unseren Zugang zum probabilistischen Modellieren – Modelle sollen dazu beitragen, die Situation des Modelleurs zu *verbessern*. Wir werden unsere Vorstellung von genuin probabilistischem Modellieren entfalten und die Idee eines Szenarios dem Modell gegenüber stellen.

5.1 Innovative Ideen

Mechanisches Rechnen und Anwenden ist im Unterricht von Wahrscheinlichkeit weit verbreitet. Wahrscheinlichkeiten oder Erwartungswerte werden aus anderen ausgerechnet, gegeben eine bestimmte Verteilung oder die Unabhängigkeit. Wahrscheinlichkeiten können mehr. Mit diesem Begriff, oder besser, mit Modellen mit Wahrscheinlichkeiten kann man aus mehreren Optionen eine als besser auszeichnen als die anderen, wenn man bestimmte Kriterien für den „Erfolg“ heranzieht.

Ein vertieftes Verständnis ist durch Modellbildung wohl leichter zu erreichen, bleibt aber noch immer schwierig genug.

Innovative Modellierungsbeispiele Wahrscheinlichkeit kann durch Modellbildung bereichert werden. Die Wahrscheinlichkeit für einen Sechser beim Würfeln ist eben nicht $1/6$. Diese Aussage begründet man dann mit der Symmetrie des Würfels und der Laplaceschen Gleichwahrscheinlichkeit. Sie wird zum Kernpunkt einer ersten Definition von Wahrscheinlichkeit. Als ob Wahrscheinlichkeit eine messbare physikalische Eigenschaft des Würfels ist. So eine enge Bindung von Wahrscheinlichkeit an das Objekt macht es später dann so schwierig, verschiedene Wahrscheinlichkeiten für ein und dasselbe Objekt miteinander zu vergleichen. Wie denn auch? Der Wahrscheinlichkeit wurde ja schon oben „ontologischer“ Charakter zugeordnet.

Warum nicht gleich von Beginn der Wahrscheinlichkeit an den Modellbildungsgedanken in den Vordergrund stellen. Kein Würfel ist symmetrisch; schon allein die Vertiefungen für die Augen zerstören die perfekte Symmetrie, von inhomogenem Material ganz zu schweigen. Entsprechend werden in Casinos die verwendeten Würfel oftmals ausgetauscht. Man kann doch von Anbeginn verschiedene Modelle vergleichen – auch für den Würfel. Wenn die Verletzung der Symmetrie gering ist, wird sich das Modell der Gleichwahrscheinlichkeit durchaus bewähren.

Also: Warum nicht die Wahrscheinlichkeit für den Sechser mit $1/6$ *modellieren* und Gründe angeben, warum man dieses Modell verwendet? Es mag zu lange dauern, bis man die tatsächlichen Abweichungen erkennen kann; die tatsächlichen Wahrscheinlichkeiten mögen nur wenig anders sein, so dass sie sich auf unsere Spielentscheidungen kaum auswirken würden.

Mit einer solchen Einstellung zu Wahrscheinlichkeiten kann auch leichter damit fertig werden, wenn Wahrscheinlichkeiten als *fiktive* Zahlen verwendet werden, die keinerlei Anspruch erheben können, einem wirklichen Wert der Wahrscheinlichkeit nahe zu kommen. So etwas trifft

speziell für Zuverlässigkeiten zu, welche oftmals durch Zahlen von ganz kleiner Größenordnung wie 10^{-10} oder noch viel kleiner repräsentiert werden. Bei der Messung und Beurteilung von Risiken ist es dasselbe. Wer hat jemals geeignete Daten vorgelegt, um die Prävalenz (die Verbreitung) von BSE zu messen? Oder, hat eine bestimmte Zahl als Wahrscheinlichkeit, dass eine Frau Brustkrebs hat, einen Sinn für genau die soeben untersuchte Frau?

Der Zweck der Modellbildung – auch mit Wahrscheinlichkeiten – ist es, in einer bestimmten Situation eine bessere Entscheidung zu treffen. Die Kriterien, welche es zu optimieren gilt, mögen Kosten sein, oder erwartete Lebensdauer, erwartete Wartezeit in einem System oder die Zuverlässigkeit eines technischen Geräts für einen bestimmten Einsatz.

Wesentlich für Modellbildung ist, dass man aus einer Bandbreite von Handlungen (oder Eingriffen) wählen kann. Das verwendete Modell mag vielleicht nicht perfekt passen – welches Modell passt denn schon völlig – aber man kann wenigstens prüfen, ob das Modell wesentliche *Eigenheiten* der Situation erfasst.

Wenn eine Lösung innerhalb des Modells gefunden wird, kann man nach kritischen Einflussgrößen (Parametern) suchen, welche die als optimal gekennzeichnete Entscheidung sehr stark beeinflussen. Man kann auch die Auswirkung einer Verletzung wesentlicher Annahmen untersuchen. Was ist, wenn eine andere Verteilung das Phänomen beschreibt? Was ist, wenn die Unabhängigkeit verletzt ist? Man kann dadurch wirklich kritische Einflussgrößen herausfiltern, über die man in der konkreten Situation mehr Information beschaffen muss oder die man besser steuern sollte etc.

Innovatives Modellbilden soll entsprechend zeigen, wie die Einbindung von Modellen auf der Basis von Wahrscheinlichkeiten helfen kann, eine Entscheidung transparenter zu treffen und nach bestimmten Kriterien zu optimieren. Beispiele dafür finden sich in Borovcnik und Kapadia (2011).

Szenarios anstelle von Modellen Die übliche Auffassung von Modell ist, dass dieses ein Bild der realen Situation darstellt, welches die Zusam-

menhänge zwischen den Objekten möglichst getreu wiedergeben soll. Und man kann das Modell verfeinern, damit es besser passt. Damit verbunden ist eine Skala, auf der man die Abweichung von Modell und realer Situation beurteilen kann. Das fehlt in der Wahrscheinlichkeitstheorie aus vielen Gründen.

Erstens basieren die Methoden zur Beurteilung, wie gut ein Wahrscheinlichkeitsmodell passt, wieder auf Wahrscheinlichkeiten. Sowohl Tests als auch Vertrauensintervalle werden durch Kennziffern beschrieben, welche bedingte Wahrscheinlichkeiten sind. Zusätzlich gelten sie nur unter der Annahme einer speziellen Verteilung (mit Ausnahme der verteilungsfreien Verfahren) und erfordern unabhängige Stichproben. Bei der Prüfung der Voraussetzung einer Verteilung befindet man sich buchstäblich in einer Falle: man hätte nämlich eine Nullhypothese zu bestätigen, was aber aus methodologischen Gründen eigentlich ausgeschlossen ist. Poppers (1935) die Idee der Bewährung von Hypothesen durch wiederholtes Testen erhärtet deren Glaubwürdigkeit, wenn sie in Replikationsstudien *nicht* abgelehnt werden können. Diese Glaubwürdigkeit ist ein *qualitatives* Argument und kann wohl durch *eine* Studie (Test) nicht wirklich hoch sein.

Zweitens muss der Test unter völlig identischen Bedingungen unabhängig wiederholt werden können, damit die bedingten Wahrscheinlichkeiten von Fehler 1. und 2. Art als relative Häufigkeiten deutbar werden. Die übliche Deutung der wesentlichen Eigenschaften von statistischen Tests entpuppt sich damit als *façon de parler* für wirkliche Anwendungen. Damit wird die Prüfung des Typs einer Verteilung durch einen Test mit relativen Häufigkeiten obsolet. In der subjektivistischen Interpretation von Wahrscheinlichkeit als Grad des Vertrauens können Modelle primär durch Plausibilitätsbetrachtungen überprüft werden – genau das machen Forscher eigentlich auch, wenn sie einen statistischen Test anwenden.

Drittens sind viele Situationen genuin singulär: *Einmal-Entscheidungen*. Wenn sie es nicht sind,

so fassen zumindest die Akteure sie als Einzelentscheidungen auf. Damit sind unabhängig wiederholbare Experimente kein geeignetes Modell für die Situation.

Viertens gibt es alternative – auch nicht-probabilistische – Modelle, die etwa in Form von Differentialgleichungen formuliert werden mögen. Es gibt keine natürliche Priorität für stochastische Modelle.

Diese Überlegungen veranlassen uns, stochastische Modelle als Szenarios aufzufassen, welche eine Situation auf einer „was wäre, wenn“-Basis erforschen lassen. Man spielt das Szenario durch und analysiert die Folgen. Obwohl ein Szenario keine perfekte Beschreibung liefert, kann man Einsichten gewinnen und Entscheidungen begründen, die, wenn schon nicht optimal, so doch transparent werden. Man kann durch Sensitivitätsanalysen von Parametern kritische Größen für die Entscheidung herausfiltern. Eigenheiten von Szenarios und Beispiele findet man in Borovcnik (2009), Borovcnik (2006) oder Borovcnik und Kapadia (2011).

Die Häufigkeiten bestimmter Unfälle sind für eine Versicherung eine Basis für die Berechnung einer Police. Für den einzelnen Autofahrer sind dagegen die persönlichen Wahrscheinlichkeiten für einen Unfall entscheidend, ob er eine Kaskoversicherung abschließt oder nicht. Innerhalb von Szenarien kann man einen Bruchpunkt finden, damit man die schwierige Schätzung der Unfallwahrscheinlichkeit umgehen kann. Liegt man darunter, keine Versicherung, darüber ist es besser, eine abzuschließen.

Ein anderes Beispiel hat mit Zuverlässigkeit zu tun. Die Annahme der Unabhängigkeit des Ausfalls einzelner Komponenten in einem technischen System ist zweifelhaft und die Zahl, die man als Zuverlässigkeit ausweist, entspringt qualitativem Wissen von Ingenieuren. Obwohl das Szenario schlecht auf ein System passt, mag man etwa daraus ablesen, ob es besser ist, einzelne Komponenten (und welche) zu verdoppeln oder zwei Systeme aufzubauen.

5.2 Ideen, die Modellbildern fördern und unterstützen

Für Modellbildern ist schematisches Anwenden zu wenig. Man muss verschiedenste Modelle miteinander vergleichen. Man muss daher nach Wegen suchen, die Mathematik indirekt zu erschließen, damit man die Modelle entsprechend versteht.

Fundamentale Idee hinter Verteilungen Es geht um die Zusammenfassung der Voraussetzungen einer Wahrscheinlichkeitsverteilung in einer tragfähigen Idee, einer fundamentalen Idee, welche diese Verteilung (auch) verkörpert. Damit erhält man einen Einblick in die interne Struktur einer Verteilung, welche bei der Modellierung auch auf die reale Situation übertragen wird. Wenn diese Eigenschaften schlecht passen, wird man die Verteilung aus der anstehenden Modellierung weglassen.

Die Normalverteilung ist mit folgender Idee verbunden: Jede Verteilung, welche sich tatsächlich oder gedacht als Summe von Zufallsvariablen ergibt, kann durch eine Normalverteilung approximiert werden. Dies ist durch Verallgemeinerungen zum zentralen Grenzwertungssatz gesichert.

Da Binomialverteilungen durch eine Summe von 0-1-Experimenten aufgebaut sind, kann man sie durch eine Normalverteilung approximieren, wenn die Zahl der Summanden groß genug ist. Auch Chi-Quadrat-Verteilungen folgen demselben Schema und sind entsprechend approximativ normal. Variable, welche aus einer zufälligen Stichprobe berechnet werden, um unbekannte Parameter zu schätzen, werden in natürlicher Weise als Summe über alle Einheiten der Stichprobe festgelegt, sodass es nicht verwunderlich ist, dass auch sie approximativ einer Normalverteilung folgen.

Andere Variable werden wiederum fiktiv auf die additive Überlagerung von vielen Einflüssen zurückgeführt. Diese Idee lag der Fehlertheorie zugrunde, wo man die Fehler von physikalischen Messungen auf die Überlagerung von Elementarfehlern zurückgeführt hat. Sogar biometrische Variable hat man sich so vorgestellt, dass sie ihren Wert durch viele andere Einflussgrößen, die

einander additiv überlagern, erhalten. Man hat diese Hypothese dann auch durch globale Messungen überprüft. So ist der Mythos der Normalverteilung entstanden.

Ein weiteres Beispiel betrifft den Poisson-Prozess, welcher ein Bindeglied zwischen Poisson- und Exponentialverteilung bildet. Die Zusammenhänge sind ähnlich wie bei der Bernoulli-Kette, der unabhängigen Wiederholung eines 0-1-Experiments, wo die Anzahl der Erfolge bei n Versuchen einer Binomialverteilung, die Wartezeit auf den nächsten Erfolg einer geometrischen Verteilung folgt. Wie bei der geometrischen Verteilung gibt es auch bei der Exponentialverteilung die Eigenschaft der Gedächtnislosigkeit: danach ist die weitere Wartezeit unabhängig davon, wie lange man schon wartet.

Die mathematischen Details findet man etwa in Meyer (1970). Borovcnik und Kapadia (2011) haben an fundamentalen Ideen hinter Verteilungen gearbeitet. Eine solche Idee fasst sozusagen die Essenz der Eigenschaften einer Verteilung zusammen, welche sich aus ihren Voraussetzungen ergibt. Das erleichtert entsprechend die konkrete Modellbildung und bietet einen Schlüssel für das Verständnis von Situationen, welche mit dieser Verteilung modelliert werden.

Probabilistische Begriffe illustrieren Es gibt einige Ansätze, mathematische Begriffe besser darzustellen und sie enger mit der Art, wie Leute denken, zu verknüpfen, was natürlich auch ihre Verwendung im Modellbildungsprozess erleichtert. Gigerenzer (2002) bringt etwa Anzahlen statt Wahrscheinlichkeiten ins Spiel und sucht nach leicht fasslichen Verkörperungen von Begriffen. Lysø (2008) setzt bei der offenen Diskussion über persönliche Vorstellungen im Anfangsunterricht an der Universität an. Borovcnik und Peard (1996) schlagen Umformulierungen vor, welche – insbesondere im Zusammenhang mit der Bayes-Formel – besser an den Zweck angepasst sind. Fischbein (1975) würde die Vorhaben als Motor für sein Wechselspiel zwischen primären und sekundären Intuitionen auffassen, welches für ihn Verstehen charakterisiert.

Es geht nicht um mehr reine Mathematik zu lernen, damit man dann später in der Modellbildung besser weiter kommt. Sowieso sind die Curricula schon überfrachtet. Beim Modellbilden muss man sich ohnehin auf eine allgemeinere Art, mit mathematischen Modellen umzugehen, einstellen. Zu viel Mathematik würde man sonst brauchen, um Modellbildung sinnvoll zu gestalten. Es muss also Wege geben, die mathematischen Inhalte über andere Medien zu erschließen.

Für die Wahrscheinlichkeitsrechnung gibt es zwei Wege dazu: Simulation zeigt die Auswirkung von Hypothesen, indem man artifizielle Daten erzeugt, welche man dann analysieren kann. Visuelle Animation mag invariante Eigenheiten dynamisch erkennen lassen oder den Einfluss verschiedener Faktoren aufzeigen. Beispiele dazu findet man in Borovcnik (2011a).

Sie umfassen die üblichen zentralen Themen wie Gesetze der großen Zahlen mit der Divergenz der absoluten und der Konvergenz der relativen Häufigkeiten. Dabei werden Schnappschüsse nach bestimmten Zeiten (Stichprobenumfängen) gemacht und für diese Zeitpunkte wird die *Verteilung* der Häufigkeiten bei Durchführung vieler Serien untersucht. Es zeigt sich, dass die Verteilung der relativen Häufigkeiten bei 100 Experimenten um rund $\pm 10\%$ schwankt, während sich bei 1000 Versuchen diese Variation of 3% reduziert. Die Konvergenz ergibt sich dann durch „Extrapolation“.

Auch der zentrale Grenzverteilungssatz wird dort abgehandelt. Auch hier werden die wesentlichen Eigenheiten in Animationen gezeigt und die mathematischen Details ausgespart. So ist die Geschwindigkeit, mit der sich die Normalisierung ergibt, abhängig von der Symmetrie der Ausgangsverteilung sowie davon, wie „kompakt oder zerissen“ die Ausgangsverteilung ist. Beim Würfeln ist die Verteilung der Mittelwerte von 20 Würfeln schon hervorragend durch eine Normalverteilung approximierbar.

6 Modellbilden – wozu die Begriffe nützlich sind

In diesem Beitrag haben wir das Feld psychologischer Forschung ausgenutzt, um wichtige Aspekte des Prozesses empirischer Forschung darzustellen. Wir sprechen abschließend die eigenartige Dualität von Gesetzmäßigkeiten und dem Zufall an, welche in empirischer Forschung zur tragenden Säule für Theorieentwicklung wird. Zu guter Letzt folgen noch ein paar Gedanken zur formativen Kraft von Modellbildung.

Magische 7 und andere Gesetze Psychologische Experimente sind eine wertvolle Quelle für den Unterricht. Aus diesen gewinnt man Beschreibungen des menschlichen Verhaltens, welche den Anspruch von allgemeinen Gesetzen haben. Die Spannung und das Interesse, das durch solche Gesetze ausgelöst wird, ist enorm.

Psychologen haben in ihren Experimenten zum Kurzzeitgedächtnis auch Musik (einzelne Noten), binäre Ziffern oder Buchstaben verwendet. Immer wieder taucht die 7 auf. Ist es wirklich wahr, dass wir uns nur 7 plus oder minus 2 Einheiten merken können? Telefonnummern bestehen üblicherweise aus 7 Ziffern, die abendländische Musik unterteilt die Tonleiter in 7 Intervalle, sogar die Likert-Skala zur Messung von Einstellungen verwendet (normalerweise) 7 Punkte. Gibt es ein archaisches Gesetz, das unser Potential, Items zu unterscheiden und sie im Gedächtnis zu behalten, einschränkt?

Andere Gesetzmäßigkeiten sprechen von 4 Informationseinheiten (Bachelder 2001); das mag sich auch in der Trennung von Telefonnummern in Gruppen von 4 und 3 Ziffern ausdrücken. Lernen ist immer auch damit verbunden, dass man Zusammenhänge (fiktive oder verborgene) „erfindet“, um die Beschränkungen unseres Gedächtnisses zu umgehen. So etwa haben Spieler im weit verbreiteten Memory-Spiel einen Vorteil, wenn sie sich eine spannende Geschichte zwischen den Bildern auf den Karten ausdenken können.

Dualität von Gesetzen und Zufall Was macht eine Gesetzmäßigkeit aus? Gesetze werden in Abgrenzung zum Zufall anerkannt. Der Zufall verkörpert dabei eine Nullhypothese, d. h., kein Effekt, keine Präferenz etc. Wenn die Daten signifikante Abweichungen davon anzeigen, dann spricht man von der „Absicherung“ einer Gesetzmäßigkeit. Diese Dualität zeigt authentisch, wie der Forschungsprozess evidenzbasiertes Wissen anhäuft.

Formative Kraft Blum (2012, S. 29) führt folgende Eigenheiten an, welche die Einbindung von Anwendungen und Modellbildung im Unterricht rechtfertigen:

- „Pragmatisch“: um Situationen in der realen Welt zu verstehen, muss man Anwendungen und Modellbildung explizit behandeln.
- „Formativ“: Kompetenzen können auch durch Modellbildungsaktivitäten verbessert werden.
- „Kulturell“: Beziehungen zur realen Welt geben erst ein adäquates Bild der Mathematik.
- „Psychologisch“: Beispiele aus der realen Welt heben das Interesse und motivieren und strukturieren mathematischen Inhalt.

Modellbildung und Simulation und die Interaktion zwischen diesen ist zentrales Thema in der didaktischen Diskussion, etwa in Chaput et al. (2011). Blum (2012) bringt auch die formative (begriffsbildende) Kraft von Modellbildung ins Spiel.

Modellbildung dient als Bindeglied zwischen Lernenden, der realen Situation und der Mathematik. Man versteht den Prozess der Modellbildung besser, die realen Situationen werden strukturiert, die mathematischen Begriffe bekommen einen Sinn, weil sie zu etwas *nützlich* sind. Das erfüllt Mathematik mit Leben. Wissenschaftliche Begriffe sind ja immer zu gewissen Zwecken entwickelt worden. Diese zu kennen, hilft, die Begriffe selbst zu verstehen.

Literatur

- Bachelder, B.L.: The magical number $4 = 7$: Span theory on capacity limitations. *Behavioral and Brain Sciences* **24**, 116–117 (2001)
- Blum, W.: *Quality teaching of mathematical modelling – what do we know, what can we do?* Plenary lecture at ICME 12. Seoul (2012) (Private Mitteilung; wird erst veröffentlicht)
- Borovenik, M.: Probabilistic and statistical thinking. In: Bosch, M. (Hrsg.) *European Research in Mathematics Education*, Bd. IV, S. 484–506. IQS Fundemi, Barcelona (2006). Online: <http://ermeweb.free.fr/CERME4/>
- Borovenik, M.: Aufgaben in der Stochastik – Chancen jenseits von Motivation. *Didaktik-Reihe der Österreichischen Mathematischen Gesellschaft* **42**, 1–23 (2009)
- Borovenik, M.: Key properties and central theorems in probability and statistics – corroborated by simulations and animations. *Selçuk Journal of Applied Mathematics*. Special Issue of „Statistics“ **12**, 3–19 (2011a)
- Borovenik, M.: Strengthening the role of probability within statistics curricula. In: Batanero, C., Burrill, G., Reading, C. (Hrsg.) *Teaching statistics in school mathematics. Challenges for teaching and teacher education: A joint ICMI/IASE Study*, S. 71–83. Springer, New York (2011b)
- Borovenik, M., Kapadia, R.: 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*, S. 1–44. Sense, Rotterdam (2011)
- Borovenik, M., Kapadia, R.: Applications of Probability: The Limerick experiments. Topic Study Group 17 ‘Mathematical applications and modelling in the teaching and learning of mathematics’ ICME 12, Seoul. (2012). Online: www.icme12.org/sub/tsg/tsg_last_view.asp?tsg_param=17
- Borovenik, M., Peard, R.: Probability. In: Bishop, A., Clements, K., Keitel, C., Kilpatrick, J., Laborde, C. (Hrsg.) *International Handbook of Mathematics Education*, S. 239–288. Kluwer, Dordrecht (1996)
- Chaput, B., Girard, J.C., Henry, M.: Modeling and simulations in statistics education. In: Batanero, C., Burrill, G., Reading, C. (Hrsg.) *Teaching statistics in school mathematics. Challenges for teaching and teacher education: A joint ICMI/IASE Study*, S. 85–95. Springer, New York (2011)
- Dubben, H.-H., Beck-Bornholdt, H.-P.: *Mit an Sicherheit grenzender Wahrscheinlichkeit. Logisches Denken und Zufall*. Rowohlt, Reinbek bei Hamburg (2010)
- Ehrenberg, A.S.C.: *Data reduction*. Wiley, New York (1981)

- de Finetti, B.: Theory of probability. Wiley, New York (1974). (Transl. A. Machi, & A. Smith)
- Fischbein, E.: The intuitive sources of probabilistic thinking in children. D. Reidel, Dordrecht (1975)
- Freedman, D., Pisani, R., Purves, S.: Statistics, 4. Aufl. Norton, London (2007)
- Gigerenzer, G.: Calculated risks: How to know when numbers deceive you. Simon & Schuster, New York (2002)
- Kataoka, V.Y., et al.: Probability teaching in basic education in Brazil: assessment and intervention. *Eleventh International Congress on Mathematics Education, TSG 13 "Research and development in the teaching and learning of probability"*, Monterrey, México. (2009). Online: http://iase-web.org/Conference_Proceedings.php?p=ICME_11_2008
- Lysø, K.: Strengths and limitations of informal conceptions in introductory probability courses for future lower secondary teachers. *Eleventh International Congress on Mathematics Education, TSG 13 "Research and development in the teaching and learning of probability"*, Monterrey, México. (2008). Online: http://iase-web.org/Conference_Proceedings.php?p=ICME_11_2008
- MacKenzie, D.A.: Statistics in Britain – 1865–1930 – The social construction of scientific knowledge. Edinburgh University Press, Edinburgh (1981)
- Meyer, P.L.: Introductory probability and statistical applications, 2. Aufl. Addison-Wesley, Reading (1970)
- Miller, G.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review* **63**(2), 81–97 (1956). Online: www.musanim.com/miller1956
- Popper, K.: *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Wien: J. Springer. 11. Aufl. H. Keuth (Hrsg.): 2005. Tübingen: Mohr Siebeck (1935)
- Richardson, M., Reischman, D.: The magical number 7. *Teaching Statistics* **33**(1), 17–19 (2011)
- Styer, D.F.: The strange world of quantum mechanics. Cambridge University Press, Cambridge (2000)
- Wild, C.J., Pfannkuch, M.: Statistical thinking in empirical enquiry. *International Statistical Review* **67**(3), 223–265 (1997)

Neue Materialien für einen realitätsbezogenen
Mathematikunterricht 2

ISTRON-Schriftenreihe

Maaß, J.; Siller, H.-S. (Hrsg.)

2014, XII, 130 S. 42 Abb., Softcover

ISBN: 978-3-658-05002-3