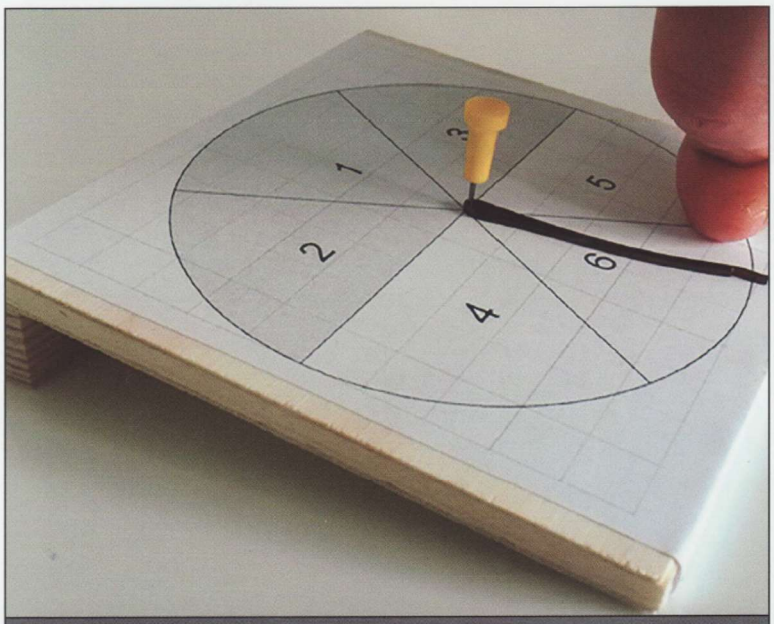
Bestell-Nr. 524210

# MU

# DER MATHEMATIK-UNTERRICHT




Wie ändern sich die Wahrscheinlichkeiten durch Schiefstellen des Glücksrades?

Stochastik

FRIEDRICH

*Jahrgang 65 · Heft 6 · Dezember 2019*

Jahrgang 65 · Heft 6–2019

MU

# DER MATHEMATIK-UNTERRICHT

Beiträge zu seiner fachlichen und fachdidaktischen Gestaltung

**Schwerpunkt** **Stochastik**
StD Dr. Wolfgang Riemer

## Adressen der Autoren

Daniel Behrens
daniel.behrens87@gmail.com

Georg Berschneider
georg.berschneider@ovgu.de

Norbert Henze
Henze@kit.edu

Henning Körner
hen.koerner@t-online.de

Wolfgang Riemer
w.riemer@arcor.de

Judith Schilling
judith.schilling@kit.edu

René L. Schilling
rene.schilling@tu-dresden.de

Reinhard Schmidt
schmidt@seminargyge.de

Günter Seebach
guenter.seebach@t-online.de

## BEITRÄGE

WOLFGANG RIEMER

# Grundvorstellungen beurteilender Statistik

## Wahrscheinlichkeit als bezweifelbares Modell
## Zweifel sortieren – Testgrößen erfinden – Testgrößen erproben

## 1 Prolog

Schülerinnen und Schüler lernen bei der Abiturvorbereitung, Rezepte abzuarbeiten, um meist aus dem Hut gezauberte Hypothesen auf einem bestimmten Signifikanzniveau verwerfen zu können. Dabei entwickeln sich keinerlei Vorstellungen darüber, was „signifikante" Testergebnisse inhaltlich bedeuten und was sie von „relevanten" unterscheidet.

Kompetenzerwerb wird zur Fassade.

Das ist insbesondere deswegen bedauerlich, weil Schüler wertvolle Vorerfahrungen zur beurteilenden Statistik aus dem Alltag mitbringen. Wenn man diese Alltagsvorstellungen ernst nimmt, wenn man Schülern etwas zutraut, ihnen die eigenen Primärintuitionen bewusst macht, sie Testgrößen selbst erfinden, erproben und dabei erfahren lässt, wie diese Größen „Bauchgefühl" präzisieren und Zweifel modellieren, entwickeln sich tragfähige Grundvorstellungen zu Hypothesentests schon in der Sekundarstufe I.

Dabei sind Experimente unerlässlich, denn tragfähige Vorstellungen entwickeln sich ausschließlich über konkrete Beispiele und gesammelte Erfahrungen [SACHS 1999, S. 10].

Den Ausführungen liegt folgendes Paradigma (ein Perspektivwechsel) zu Grunde: Wahrscheinlichkeiten (auch die LAPLACEschen!) werden ab Klasse 7 nicht als objektive Größen, sondern als vom Menschen gesetzte Modelle verstanden, die die Wirklichkeit nie genau, sondern nur mehr oder weniger gut beschreiben. Und Testgrößen nutzt man nicht im Sinne eines schwarz/weiß ja/nein Denkens zum Akzeptieren/Verwerfen, sondern als Werkzeug, das hilft, *bei festem Stichprobenumfang* die Güte von Modellen miteinander zu vergleichen. Ohne fixierten Stichprobenumfang ist nämlich – zumindest für einen Anfänger – das Interpretieren von Testergebnissen unmöglich. Das liegt am „Hauptsatz" der beurteilenden Statistik

> Jede Hypothese ist falsch – und kann durch Erhöhen des Stichprobenumfanges auf jedem Signifikanzniveau verworfen werden.

Kaum ein Lehrer ist sich dieses Hauptsatzes bewusst. Schulbücher und Bildungsstandards verschweigen ihn, weil er standardisierte Testrezepte ad absurdum führt.

## 2 Das Problem

Für viele Abiturienten bleibt das Testen von Hypothesen ein „Buch mit sieben Siegeln", für viele Lehrer ein angstbesetztes Thema. Meist werden ohne authentische Fragestellungen mühsam eingekleidete „Kontextaufgaben" entkleidet. Was dabei herauskommt, hat mit dem gesunden Menschenverstand meist wenig zu tun, weil man standardisierte Aufgabenformate bedienen muss, nicht den gesunden Menschenverstand. Eingängige Fehlinterpretationen sind in einer Vielzahl empirischer Untersuchungen belegt, vgl. KRAUSS & WASSNER [2000], DIEPGEN [2002].

Die möglicherweise gängigste wird selbst in vielen kommerziellen „Erklärvideos“ verbreitet: „Ich konnte die Hypothese $H_0$ mit der Irrtumswahrscheinlichkeit 5 % verwerfen, also gilt die Alternative $H_1$ mit 95 % Sicherheit, denn $H_0$ und $H_1$ sind Gegenereignisse.“

Das kommt daher, dass wir im Alltag (**Abb. 1a**) ständig subjektive Wahrscheinlichkeiten nutzen, um auszudrücken, wie sehr wir an die Gültigkeit von Hypothesen glauben. Und Lernende erwarten von beurteilender Statistik eine Quantifizierung genau dieser Unsicherheit. Aber beurteilende Statistik liefert keine Wahrscheinlichkeiten für die Gültigkeit von Hypothesen, sondern nur Wahrscheinlichkeiten für (kritische) Ereignisse bei unterstellter Gültigkeit von Hypothesen, die eigentlich niemanden interessieren.


**London.** Der britische Handelsminister Liam Fox hat die Wahrscheinlichkeit für einen Brexit ohne Abkommen („No Deal“) auf 60 Prozent geschätzt. Verantwortlich dafür sei die „Unnachgiebigkeit“ Brüssels und vor allem des EU-Unterhändlers Michel Barnier, kritisierte Fox in der „Sunday Times“. (afp)

*Minister Liam Fox Foto: dpa*

**Abb. 1a**: subjektive Wahrscheinlichkeit KSTA 06. 08. 2018

Als Konsequenz fordert BUTH [2003, S. 30]: „Also gehören Signifikanztests in den Papierkorb und zwar auf allen drei Ebenen: in der Mathematik, in den anwendenden Disziplinen und im Mathematikunterricht.“

DIEPGEN [2002 S. 34] konstatiert: „Meinen Optimismus habe ich längst verloren. …

Ich habe das Vermögen in den Humanwissenschaften und im Schulunterricht unterschätzt, auch an den sinnlosesten Dingen über Jahrzehnte festzuhalten.“

Ein weiteres Jahrzehnt später formuliert MOßBURGER [2014, S. 7]: „Wozu Signifikanztests? Was weiß man über die Gültigkeit einer Hypothese, wenn eine Stichprobe im Ablehnungsbereich A liegt? Nichts. Was folgt aus $P_H(A) \leq 5\,\%$? Nichts. Wozu soll man dann $P_H(A)$ berechnen? Um eine Meinung über eine Hypothese zu begründen, die mathematisch aber nicht begründet werden kann? Sind Signifikanztests in der Schule überhaupt sinnvoll?

Ich fürchte, Aussagen wie ‚Ich lehne H auf einem Signifikanzniveau von 5 % ab‘ werden in unserer Gesellschaft immer wieder dazu benutzt, mathematisch fundierte Erkenntnisse vorzugaukeln. Solange Zahlen für pseudo-mathematische Argumente missbraucht werden, solange sollte der Mathematikunterricht darüber aufklären, wie wenig ein Signifikanztest aussagt.“


Schickt die statistische Signifikanz in den Ruhestand!

Drei Statistiker fordern gemeinsam mit mehr als 800 weiteren Fachleuten, den p-Wert als Signifikanzkriterium aufzugeben: Er unterstelle zwei Kategorien von Ergebnissen, die es eigentlich nicht gibt.

Valentin Amrhein, Sander Greenland und Blake McShane

**Abb. 1b**: Spektrum der Wissenschaft, 30. 05. 2019

Noch aktueller und von erheblicher Breitenwirkung ist der Aufruf AMRHEIN [2019], der von 800 Statistikern unterzeichnet und in einer Übersetzung von „Spektrum der Wissenschaft“ wie in **Abb. 1b** getitelt wurde.

Lehrer, die ihre Schüler seit Jahren zu erfolgreichen Abiturienten ausbilden, mögen diese kritischen Positionen für überzogen halten. Aber auch sie geraten bei der Suche nach Antworten auf folgende Fragen 2.1 und 2.2 eines pfiffigen Schülers „Max“ schnell ins Grübeln.

Liegt das vielleicht doch daran, dass wir über den Sinn des Hypothesentestens im Unterricht viel zu wenig nachdenken?

**2.1** „Der Hypothesentest ist ein rechnerisches Verfahren, das eine Entscheidungshilfe gibt, ob ein Würfel ideal ist oder nicht.“ [https://lehrerfortbildung-bw.de/umatnatech/mathematik/gym/bp2004/fb2/modul4/4_unterricht/] (am 04.11.2019 gelöscht).

Max: „Aber wir wissen doch, dass es keine idealen Würfel gibt, genauso wenig wie echte Punkte in der Geometrie, dazu brauche ich keine Entscheidungshilfe.“

**2.2** Max: a) „Ich konnte die Hypothese $H_0$: $p = 0{,}5$ bei $n = 100$ auf dem 5 % Signifikanzniveau verwerfen. Aber was bedeutet das jetzt?“

b) Wie ändert sich die Bedeutung, wenn dem Verwerfen der Stichprobenumfang nicht $n = 100$, sondern $n = 25$ oder $n = 400$ zugrunde lag?

c) Wie ändert sich die Bedeutung, wenn ich nicht auf dem 5 %- sondern auf dem 1 %-Signifikanzniveau verwerfen konnte? Welche Stichworte muss ich in die Merkheft-Tabelle eintragen, um die Bedeutungsunterschiede prägnant darzustellen?“

**Tabelle 1**:

| | n = 25 | n = 100 | n = 400 |
|---|---|---|---|
| α = 5 % | | | |
| α = 1 % | | | |

## 3 Zwei „klassische“ Lösungen

In der Literatur finden sich zwei Auswege aus dem Dilemma zwischen der weiten Verbreitung von Signifikanztests einerseits – und der Fragwürdigkeit ihrer schematischen Anwendung andererseits.

### 3.1 Lösung 1: Schätzen statt Testen – der absolute Favorit

Der nahe liegende – mit den Bildungsstandards kompatible – Ausweg besteht darin, unbekannte Wahrscheinlichkeiten durch Konfidenzintervalle zu schätzen, statt Hypothesen über irgendwelche Wahrscheinlichkeiten zu testen. Tatsächlich sind Konfidenzintervalle natürlicher und vernetzender als Hypothesentests, denn sie übertragen die fundamentale Idee des Messens und der Messungenauigkeiten von der Geometrie in die Stochastik. Man deutet die Länge des Konfidenzintervalls als Messungenauigkeit und erkennt, dass eine Vervierfachung (!) des Versuchsumfangs die Messungenauigkeit halbiert ($1/\sqrt{n}$ -Gesetz). Eine Vergrößerung des Stichprobenumfanges entspricht damit – in die Geometrie übertragen – dem Blick durch ein Vergrößerungsglas beim Ablesen von Längenskalen.

Bei einseitigen Hypothesentests kommt extrem erschwerend hinzu, dass statt einer Suche nach der Wahrheit das Aufspüren oft künstlich versteckter Interessenlagen im Vordergrund zu stehen scheint. In der Tat bringt die Suche nach der Nullhypothese oder die Frage, ob man rechts- oder linksseitig testen soll, so manche Lerngruppe und auch deren Lehrer zur Verzweiflung, vgl. STOYAN [2011] und die dort abgedruckte Abituraufgabe (NRW 2008). Die „Reinlich und Sohn“-Aufgabe hat inzwischen einen durchaus zweifelhaften „Kultstatus“ als Urmutter aller Signifikanztest-Abituraufgaben.

Leider gibt es viele Bundesländer (Nordrhein-Westfalen, Baden-Württemberg, Bayern), die Schülern den genetischen Weg, über Konfidenzintervalle in die beurteilende Statistik einzusteigen, vorenthalten, obwohl es ausgezeichnete Schulbuchvorlagen und ausgearbeitete Handreichungen gibt: LERGENMÜLLER [2012, S. 151 ff.], BRANDT [2018, S. 122 ff.] und RIEMER [https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=3498].

Der Grund dafür ist vermutlich, dass man – anders als bei Konfidenzintervallen – bei Hypothesentests mit Binomialtabellen (mit oder ohne GTR) auskommt und auf die Sigmaregeln verzichten kann. Die hinter den Sigmaregeln steckende Normalverteilung (den Satz von de MOIVRE-LAPLACE) braucht man dann nicht zu erwähnen.

### 3.2 Lösung 2: Die BAYESsche Sicht auf Hypothesentests

In allen Bundesländern kennen Schülerinnen und Schüler vor der Binomialverteilung Vierfeldertafeln, bedingte Wahrscheinlichkeiten und die Regel von BAYES. Wenn man ihnen in

diesem Rahmen erlaubt, der Primärintuition zu folgen und die Gültigkeit von Hypothesen durch priori-Wahrscheinlichkeiten zu bewerten, kann man studieren – und in Spielsituationen auch eindrucksvoll erleben – wie sich diese Wahrscheinlichkeiten durch Beobachtung kritischer Testwerte verändern: In der Regel werden durch priori-Wahrscheinlichkeiten bewertete Hypothesen tatsächlich unwahrscheinlicher, wenn man sie auf dem 5 %-Signifikanzniveau verwerfen kann, vgl. RIEMER [1986], KRAUS & WASSNER [2000], MOSSBURGER [2014], MOTZER [2017]. Das wirft einerseits Licht auf die Logik von Signifikanztests, beugt aber andererseits der Verfestigung oben zitierter Fehlvorstellungen vor. Für Aussagen über die Wahrscheinlichkeit der Gültigkeit von Hypothesen benötigt man spezifizierte Alternativen mit priori-Bewertungen, die man in Spielsituationen leicht schaffen kann, die aber in der Praxis meist fehlen. Wenn man in solchen Spielsituationen Fehlentscheidungen durch Verluste bewertet, wird darüber hinaus auch der Unterschied zwischen Nullhypothese $H_0$ und der Alternative $H_1$ transparent – und Schüler erleben, wie Signifikanztests versuchen, Verluste zu minimieren, nicht aber die Wahrscheinlichkeit von Fehlentscheidungen, vgl. RIEMER [2020].

Der Blick durch die „BAYESsche Brille“ entspricht damit einerseits der von MOSSBURGER [2014] geforderten Aufklärung über die eingeschränkte Bedeutung von Signifikanztest.

Er berücksichtigt andererseits auch die Tatsache, dass Verbote und Belehrungen weniger wirksam sind als selbstständig experimentell zu erkunden, welche Aussagen unter welchen Umständen möglich sind. Die „BAYESsche Lösung“ aus dem Dilemma ist also in erster Linie eine pädagogische, die bekanntes Wissen mit neuem vernetzt. Es gibt jedoch derzeit keinen Lehrplan, in dem sich diese Erkenntnis widerspiegelt.

## 4 Wahrscheinlichkeiten als Modelle und das Konzept des Bezweifelns

Im Folgenden wird als Alternative zu 3.1 und 3.2 ein ganzheitlicher Einstieg in das Testen von Hypothesen beschreiben. Er ist so elementar und intuitiv naheliegend, dass er Schüler ab Klasse 8 schon im Rahmen der Wahrscheinlichkeitsrechnung wirksam gegen die in 2 genannten Fehlvorstellungen impft. Dabei hilft es, *Wahrscheinlichkeitsverteilungen als vom Menschen gesetzte Modelle der Realität* zu deuten, die nicht wie in der Logik des Signifikanztests angenommen oder verworfen werden, sondern deren Bewertung und sukzessive Verbesserung den Modellbildungskreislauf in idealtypischer Weise erfahrbar macht. Dabei ist es sehr hilfreich, die nach SACHS [1999] unerlässlichen Erfahrungen

- mit einem festen Stichprobenumfang zu sammeln ($n = 120$ ist praktikabel) und
- gleich Wahrscheinlichkeitsverteilungen statt einzelner Wahrscheinlichkeiten zu studieren, also mit gezinkten Würfeln zu experimentieren statt nur mit gezinkten Münzen.

### 4.1 Modellbildungskreislauf – Wahrscheinlichkeit als Modell

„Klassisch“ wird Wahrscheinlichkeit als objektive, einer Situation innewohnende und eindeutige, Größe aufgefasst, die man entweder als Anteil genau berechnet oder als Grenzwert relativer Häufigkeit „bestimmt“ (dabei aber mit schlechtem Gewissen „unter den Teppich kehrt“, dass das praktisch unmöglich ist). Man spricht vom LAPLACEschen bzw. vom frequentistischen Wahrscheinlichkeitsbegriff, der eng mit dem Namen v. MISES verknüpft ist.

Tatsächlich ist Schülern, sobald sie über den Prozentbegriff verfügen, der „hypothetisch-prognostische Wahrscheinlichkeitsbegriff“, der seinerseits eng mit KOLMOGOROFF verknüpft ist, aus dem Alltag viel vertrauter. Wahrscheinlichkeiten werden in diesem Konzept erlebt als

vom Menschen gesetzte Zahlen, als Modelle also, die vorhersagen wollen, um welchen Wert relative Häufigkeiten zufallsabhängig pendeln werden, wenn man plant, einen Versuch oder eine Datenerhebung mehrfach zu wiederholen.

Der Unterschied zum LAPLACEschen und zum frequentistischen Wahrscheinlichkeitsbegriff wird beim Würfeln mit Quadern an folgendem Impuls deutlich, vgl. RIEMER [2016]:

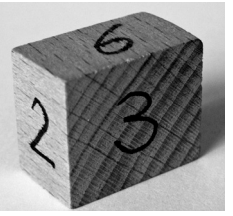
„Schaut euch die Quader genau an und schätzt die Chancen der sechs Augenzahlen in Prozent." Nicht würfeln, nur schätzen! Nach Gefühl!

Die Kinder akzeptieren das Würfelverbot und nennen bereitwillig „Prozentzahlen", von denen man einige an der Tafel festhält (**Abb. 2b**). Dabei beachten sie intuitiv

a) Gegenseiten haben gleiche Chancen.
b) Große Seiten haben große Chancen.
c) Alle Chancen addieren sich zu 100 %.

Tatsächlich haben diese Zeilen alles, was man nach KOLMOGOROFFS Axiomen[3] von Wahrscheinlichkeitsverteilungen verlangt, vor allem aber spiegeln sie Intuitionen der Schüler wider und das Gefühl: „Mathe hat hier etwas mit mir zu tun."

**a**




**b**

| Würfeln mit Würfelbechern - auf den Tisch gestülpt | | | | | | |
|---|---|---|---|---|---|---|
| Schätzungen | 1 | 2 | 3 | 4 | 5 | 6 |
| Renè | 10% | 4% | 35% | 35% | 5% | 10% |
| Stefan | 15% | 10% | 25% | 25% | 10% | 15% |
| Alexa | 10% | 12% | 35% | 20% | 15% | 8% |
| Joanna | 15% | 15% | 20% | 20% | 15% | 15% |
| Jasmin | 15% | 5% | 30% | 30% | 5% | 15% |
| Flache in cm² | 2.99 | 2.60 | 4.60 | 4.60 | 2.60 | 2.99 |
| Fläche in % | 14.7% | 12.8% | 22.6% | 22.6% | 12.8% | 14.7% |

**Abb. 2**: **a** Quader $1{,}3 \times 2 \times 2{,}3\,\text{cm}^3$ und **b** einige Schätzungen. Alexas Schätzung führt zu intensiver Diskussion. Der Unterschied zwischen unsymmetrischen Häufigkeitsverteilungen (Realitätsebene) und symmetrischen Wahrscheinlichkeiten (Modellebene) tritt prägnant hervor – deswegen sind teilsymmetrische Objekte wie die Quader ideal.

Das Bedürfnis, diese Prognosen an der Realität zu überprüfen und nach dem Motto „aus Erfahrung wird Erwartung" anschließend zu verbessern (Modellierungskreislauf), garantiert eine spannende Unterrichtsstunde, in der Schüler – anders als beim Berechnen von LAPLACE-Wahrscheinlichkeiten – nicht nur Zufallsschwankungen erleben, sondern auch erfahren, dass sie bei größeren Stichproben zwar kleiner werden, aber prinzipiell nicht auszumerzen sind.

Als Abschluss einigt man sich für zukünftige Rechnungen in der Modellebene auf brauchbare (symmetrische) Wahrscheinlichkeitsverteilungen, in die man sehr viel mehr Vertrauen hat als in die zuvor „aus dem hohlen Bauch heraus geschätzten" (letzte Zeilen von **Abb. 3**).

| | | | | | | |
|---|---|---|---|---|---|---|
| Patrick | 10 | 6 | 28 | 41 | 4 | 11 |
| Daniel | 6 | 7 | 35 | 45 | 4 | 3 |
| Binoy | 7 | 4 | 37 | 34 | 1 | 17 |
| Tobias | 3 | 6 | 48 | 33 | 6 | 4 |
| Michael | 12 | 0 | 28 | 42 | 7 | 11 |
| absolute H. | 38 | 23 | 176 | 195 | 22 | 46 |
| relative H. | 7.6% | 4.6% | 35.2% | 39.0% | 4.4% | 9.2% |
| | | | | | | |
| Paula | 11 | 6 | 34 | 32 | 7 | 10 |
| Elaine | 14 | 10 | 28 | 24 | 9 | 15 |
| Marie | 4 | 6 | 41 | 32 | 11 | 6 |
| Marga | 10 | 6 | 34 | 29 | 7 | 14 |
| Sandra | 7 | 4 | 30 | 37 | 4 | 18 |
| absolute H. | 46 | 32 | 167 | 154 | 38 | 63 |
| relative H. | 9.2% | 6.4% | 33.4% | 30.8% | 7.6% | 12.6% |
| | | | | | | |
| Summe (27 Kinder) | 279 | 207 | 834 | 883 | 204 | 293 |
| | 10.3% | 7.7% | 30.9% | 32.7% | 7.6% | 10.9% |
| | | | | | | |
| verbesserte Schätzungen | | | | | | |
| Hypothese A | 11.0% | 8.0% | 31.0% | 31.0% | 8.0% | 11.0% |
| Hypothese B | 19.5% | 8.0% | 31.5% | 31.5% | 8.0% | 10.5% |

**Abb. 3**: Häufigkeiten und verbesserte Wahrscheinlichkeiten

Die Wahrscheinlichkeiten bleiben *vom Menschen gesetzte Modelle*, die gut zur Wirklichkeit passen, aber möglicherweise durch weitere Versuche noch verbessert werden können (Modellbildungskreislauf). Das „Erlebte" lässt sich wie folgt zusammenfassen:

Wahrscheinlichkeiten sagen relative Häufigkeiten voraus. Sie sind als Modelle gut *gewählt*, wenn die zufällig schwankenden relativen Häufigkeiten um die Wahrscheinlichkeit pendeln, also etwa gleich oft über wie unter der Wahrscheinlichkeit liegen.

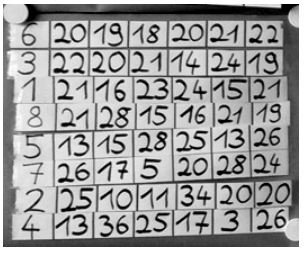
### 4.2 Den Zweifel sortieren – ein Unterrichtsgang

Genau genommen hat man nach 4.1 mit dem hypothetisch-prognostischen Wahrscheinlichkeitsbegriff und dem Modellierungskreislauf von Anfang an die Brücke zwischen Wahrscheinlichkeitsrechnung und beurteilender Statistik geschlagen. Hypothesen, die gefühlsmäßig schlechte Vorhersagen machen, werden verbessert. Man vertraut darauf, dass die verbesserten Hypothesen die Wirklichkeit zwar nicht perfekt, aber „für den Hausgebrauch" hinreichend gut beschreiben.

Aber was bedeutet: „gefühlsmäßig schlechte" Vorhersagen? Das zu untersuchen ist ab Klassenstufe 8 motivierend, weil wiederum die eigene Intuition im Mittelpunkt steht. Man konfrontiert die Schüler mit dem folgenden

**Arbeitsauftrag 1**:
In jeder Zeile von **Abb. 4a** wurde 120-mal „gewürfelt". Diskutiert in Kleingruppen, welche Zeilen gefühlsmäßig gut von einem fairen Spielwürfel stammen könnten. Sortiert nach steigendem Zweifel!

| | Augenzahlen | | | | | |
|---|---|---|---|---|---|---|
| No | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 21 | 16 | 23 | 24 | 15 | 21 |
| 2 | 25 | 10 | 11 | 34 | 20 | 20 |
| 3 | 22 | 20 | 21 | 14 | 24 | 19 |
| 4 | 13 | 36 | 25 | 17 | 3 | 26 |
| 5 | 13 | 15 | 28 | 25 | 13 | 26 |
| 6 | 20 | 19 | 18 | 20 | 21 | 22 |
| 7 | 26 | 17 | 5 | 20 | 28 | 24 |
| 8 | 21 | 28 | 15 | 16 | 21 | 19 |



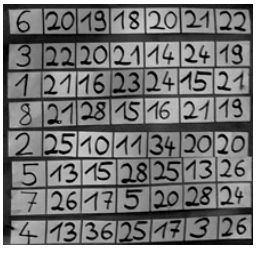




**Abb. 4**: **a** Ausgangsverteilungen **b** Sortierung Gruppe 1 **c** Sortierung Gruppe 2

Dieser Auftrag führt zu lebhaftem Spekulieren über die Größe von Zufallsschwankungen beim 120-maligen Rollen eines LAPLACE-Würfels – und einem intensiven Gedankenaustausch. Überraschend ist, wie wenig sich die Sortierungen verschiedener Gruppen voneinander unterscheiden. **Abb. 4b** und **c** zeigen zwei Beispiele.

Einen Ansatz zur Objektivierung dieser Intuition liefert Arbeitsauftrag 2.

### 4.3 Sortiergrößen erfinden

**Arbeitsauftrag 2:**
Entscheidet rechnerisch, ob die Abweichungen vom Ideal (20|20|20|20|20|20) in der Zeile 5 oder in der Zeile 6 (von **Abb. 4a**) größer sind. Erfindet dazu eine – besser noch: möglichst viele – Sortiergrößen S mit der Eigenschaft: „Je größer S, desto größer der Zweifel."

Wie **Abb. 5** zeigt, überraschen Schüler ab Klassenstufe 8 auch hier mit konstruktiven Ideen, wenn man ihnen etwas Zeit zum Nachdenken lässt. Falls quadratische Funktionen verankert sind, wird neben der Summe $s$ der betraglichen Abweichungen von den Idealwerten oft auch die Sortiergröße $q$ genannt, die durch das Quadrieren große Abweichungen stärker „bestraft" als $s$. Die Division durch den Idealwert 20 macht $q$ numerisch handlicher. Dass es sich bei $t = q/20$ um die Chiquadrat-Testgröße handelt, kann man beiläufig erwähnen.[1]

| 4 | 13 | 28 | 16 | 29 | 30 |
|---|---|---|---|---|---|

- ***g***: die ***größte*** |Abweichung| vom Sollwert 20 -> 16
- ***s:*** die ***Summe aller*** |Abweichungen| vom Sollwert -> 54
- ***q:*** die ***Summe der Quadrate aller*** Abweichungen -> 566
- ***t :*** q/20 (handlicher als q) -> 28.3
- ***u:*** der Unterschied zwischen der größten und der kleinsten Häufigkeit -> 26
- ***a:*** die Anzahl der Augenzahlen, die besonders selten (≤10) oder besonders häufig (≥ 30) auftraten -> 2

**Abb. 5**: Vorschläge für geeignete Sortiergrößen und numerische Ergebnisse für eine Beispielverteilung

| | kritische Grenze | | | | | | 11 | 30 | 11 | 18 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Σ (H-20)²/20 | Σ\|H-20\| | MAX\|H-20\| | MAX(H)-MIN(H) | #≤10+#≥30 |
| No | 1 | 2 | 3 | 4 | 5 | 6 | t | s | g | u | a |
| 1 | 21 | 16 | 23 | 24 | 15 | 21 | 120 | 3.4 | 18.0 | 5 | 9 | 0 |
| 2 | 25 | 10 | 11 | 34 | 20 | 20 | 120 | 20.1 | 38.0 | 14 | 24 | 2 |
| 3 | 22 | 20 | 21 | 14 | 24 | 19 | 120 | 2.9 | 14.0 | 6 | 10 | 0 |
| 4 | 13 | 36 | 25 | 17 | 3 | 26 | 120 | 33.2 | 54.0 | 17 | 33 | 2 |
| 5 | 13 | 15 | 28 | 25 | 13 | 26 | 120 | 12.4 | 38.0 | 8 | 15 | 0 |
| 6 | 20 | 19 | 18 | 20 | 21 | 22 | 120 | 0.5 | 6.0 | 2 | 4 | 0 |
| 7 | 26 | 17 | 5 | 20 | 28 | 24 | 120 | 17.5 | 36.0 | 15 | 23 | 1 |
| 8 | 21 | 28 | 15 | 16 | 21 | 19 | 120 | 5.4 | 20.0 | 8 | 13 | 0 |

**Abb. 6**: Sortiergrößen der Verteilungen aus **Abb. 4a**

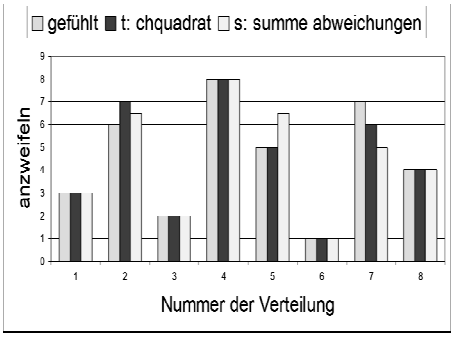



**Abb. 7**: Positionen nach Sortierung durch zu t und s im Vergleich zu einer „intuitiv gefühlten" Sortierung.

Es ist überraschend, dass die von den Schülern erfundenen Testgrößen ähnlich sortieren, vor allem aber, dass sie ausgezeichnet mit den intuitiven aus **Abb. 4b** und **4c** zusammenpassen. Wieder hat „Mathe hat etwas mit uns zu tun".

Das gilt auch für den folgenden Arbeitsauftrag 3.

### 4.4 Die Sortiergrößen werden Testgrößen

**Arbeitsauftrag 3:**

Betrachtet z. B. die Sortiergröße $s$ (Summe aller betraglichen Abweichungen) und einigt euch in eurer Gruppe auf Obergrenzen, also „Benchmarks", ab denen ihr „Zweifel" bzw. „starke Zweifel" an der Fairness des Würfels für gerechtfertigt haltet.

Naturgemäß differieren die von Gruppen genannten Benchmarks etwas. Vielfach orientiert man sich an den Verteilungen aus **Abb. 3** und meldet z. B. bei Zeile 5, also ab $s = 38$, Zweifel und bei Zeile 4, also ab $s = 54$, starke Zweifel an.

Diese unterschiedlichen Positionen sind Anlass für den folgenden

**Arbeitsauftrag 4:**

Würfelt mit echten, dann mit computersimulierten, Spielwürfeln 120-mal und sucht Obergrenzen der Sortiergröße $g$, $s$ und $t$, die nur selten (in 5 % aller Fälle) überschritten wird.

Die Verlangsamung durch das händische Würfeln und das Zusammenfassen der zugehörigen Sortiergrößen im Klassenverband, die dabei ablaufenden Gespräche in einer Atmosphäre emotionaler Eingebundenheit sollte man vor Computersimulationen Lerngruppen keinesfalls vorenthalten.

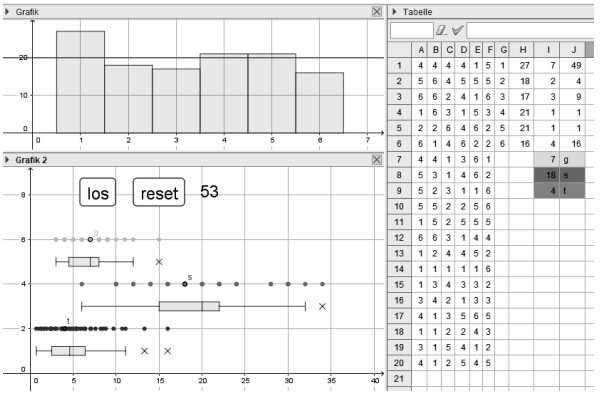



**Abb. 8**: 53 Computersimulationen des 120-maligen Würfelns. Die Verwendung von Boxplots und Ausreißern visualisiert, welche Sortiergrößen, hier g (größte Abweichung vom Sollwert 20), s (Summe aller betraglichen Abweichungen vom Sollwert) und t (Chiquadrat) für LAPLACE-Würfel ungewöhnlich sind

Die Experimente zeigen: Bei 120 Würfen eines LAPLACE-Würfels gilt fast immer (in ca. 95 % aller Fälle) $g<11$, $s<30$ und $t<11$. Dadurch, dass man in der Computersimulation (**Abb. 8**) neben den Sortiergrößen auch die zugehörige Häufigkeitsverteilung visualisiert, entwickelt sich ein sicheres Gefühl dafür, wie Verteilungen aussehen, die mit $s>30$, $g>11$ $t>11$ die kritischen Grenzen überschreiten. Durch folgende – auch in der „scientific community" nur ausgehandelte – Vereinbarung *werden die Sortiergrößen zu Testgrößen*: „Zweifel an der Fairness des Würfels sind spätestens dann gerechtfertigt, wenn die kritische Grenze überschritten werden."

Wenn man sensibel mit der Sprache umgeht und den „gefühlten Zweifel" als Begriff pflegt, kommt niemand auf die (absurde) Idee, nach der Wahrscheinlichkeit zu fragen, mit der ein verwendeter Würfel „laplacesch" ist. Es geht einzig und allein um die Einschätzung, wie gut „für den Hausgebrauch" das LAPLACE-Modell für ein verwendetes „Würfelobjekt" taugt.

### 4.5 Testgrößen erproben

Ganz im Sinne des Eingangszitats „Tragfähige Vorstellungen entwickeln sich ausschließlich über konkrete Beispiele und gesammelte Erfahrungen …" wird man sich dem Reiz, die unterschiedlichen Testgrößen zu erproben – und deren Wirksamkeit miteinander zu vergleichen –, nicht entziehen. Nicht nur wegen der unterrichtlichen Praktikabilität, sondern auch wegen der verblüffenden Ergebnisse eignen sich dazu folgende Objekte ausgezeichnet:

- beschriftete Sechskantbleistifte, die man über den Tisch rollt. Beide Rollrichtungen getrennt untersuchen!
- Glücksräder aus Haarklemmen, die man durch Schnippen um einen Pin auf einer schiefen Ebene rotieren lässt. – Neigung der Ebene variieren!
- Gummibären, die man nach den sechs Farben/Geschmacksrichtungen auszählt.

Diese Realexperimente kann man durch Computersimulationen abrunden, bei denen man die Gleichverteilungen manipuliert – wohlgemerkt abrunden, nicht ersetzen!

Forschungsfragen:

- Wie oft liefern die Testgrößen beim Stichprobenumfang $n = 120$ Anlass zum Zweifel an der Gleichwahrscheinlichkeit der sechs Ergebnisse?
- Welche der Testgrößen schüren am schnellsten Zweifel?

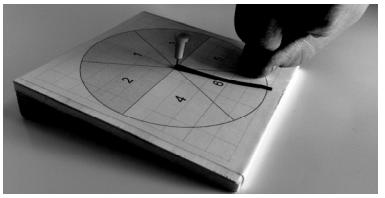


**Abb. 9**: **a** Glücksrad und **b** Ergebnistabelle. Eine Haarklemme rotiert auf einer schiefen Ebene mit Neigung 10 % im Uhrzeigersinn um einen Pin. Bei n = 120 Schnipsern signalisieren t und s in 14 von 16 Fällen Zweifel an der Gleichwahrscheinlichkeit der Felder 1 … 6. In diesen Fällen sind auch die anderen Testgrößen zumindest „erhöht". Die letzte Zeile mit den kumulierten relativen Häufigkeiten signalisiert: „Bei den bergab-Feldern 3 und 5 fehlt die Wahrscheinlichkeit, die sich bei den bergauf-Feldern 2 und 4 anlagert." Beim Wechseln der Drehrichtung wird das Glücksrad perfekt durch eine Gleichverteilung beschrieben.

| | kritische Grenze | | | | | | | 11 | 30 | 11 | 18 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Σ (H-20)²/20 | Σ\|H-20\| | MAX\|H-20\| | MAX(H)-MIN(H) | #≤10+#≥30 |
| Dominique | 17 | 28 | 15 | 32 | 5 | 23 | | t | s | g | u | a |
| Oliver | 25 | 30 | 11 | 29 | 11 | 14 | 120 | 20.2 | 48.0 | 10 | 19 | 1 |
| Hylia | 12 | 23 | 14 | 28 | 13 | 30 | 120 | 16.1 | 42.0 | 10 | 18 | 1 |
| Renate | 22 | 26 | 10 | 31 | 16 | 15 | 120 | 15.1 | 38.0 | 11 | 21 | 2 |
| Julia | 23 | 32 | 9 | 21 | 10 | 25 | 120 | 20.0 | 42.0 | 12 | 23 | 3 |
| Nora | 26 | 26 | 13 | 25 | 8 | 22 | 120 | 14.7 | 38.0 | 12 | 18 | 1 |
| Lena | 20 | 32 | 9 | 28 | 11 | 20 | 120 | 20.5 | 40.0 | 12 | 23 | 2 |
| Nadia | 16 | 39 | 8 | 31 | 9 | 17 | 120 | 38.6 | 60.0 | 19 | 31 | 4 |
| Daniel | 20 | 29 | 10 | 30 | 11 | 20 | 120 | 18.1 | 38.0 | 10 | 20 | 2 |
| Maren | 19 | 33 | 14 | 28 | 12 | 14 | 120 | 18.5 | 42.0 | 13 | 21 | 1 |
| Sabine | 21 | 27 | 11 | 19 | 12 | 30 | 120 | 14.8 | 36.0 | 10 | 19 | 1 |
| Lutz | 24 | 28 | 11 | 26 | 14 | 17 | 120 | 12.1 | 36.0 | 9 | 17 | 0 |
| Nicole | 21 | 26 | 16 | 24 | 14 | 19 | 120 | 5.3 | 22.0 | 6 | 12 | 0 |
| Tobias | 25 | 19 | 16 | 31 | 17 | 17 | 125 | 8.5 | 27.0 | 11 | 15 | 1 |
| Thomas | 20 | 24 | 11 | 33 | 8 | 24 | 120 | 21.3 | 42.0 | 13 | 25 | 2 |
| Matthias | 20 | 24 | 11 | 28 | 11 | 27 | 121 | 14.4 | 37.0 | 9 | 17 | 0 |
| | 17% | 23% | 10% | 23% | 9% | 17% | | | | | | |


**Abb. 10a**: Mit den Augenzahlen 1, …, 6 beschrifteter Sechskant-Bleistiftwürfel

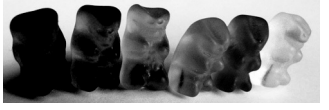


**Abb. 11a**: 1 = Himbeere, 2 = Erdbeere, 3 = Orange, 4 = Zitrone 5 = Apfel, 6 = Ananas

| | kritische Grenze | | | | | | | 11 | 30 | 11 | 18 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Σ (H-20)²/20 | Σ\|H-20\| | MAX\|H-20\| | MAX(H)-MIN(H) | #≤10+#≥30 |
| Julian re | 7 | 17 | 18 | 36 | 22 | 20 | | t | s | g | u | a |
| Julian li | 15 | 26 | 32 | 8 | 15 | 24 | 120 | 19.5 | 44.0 | 12 | 24 | 2 |
| Simon re | 27 | 21 | 19 | 31 | 18 | 4 | 120 | 21.6 | 38.0 | 16 | 27 | 2 |
| Simon li | 31 | 15 | 1 | 16 | 24 | 33 | 120 | 35.4 | 56.0 | 19 | 32 | 3 |
| Svenja re | 27 | 19 | 12 | 27 | 20 | 15 | 120 | 9.4 | 28.0 | 8 | 15 | 0 |
| Svenja li | 10 | 20 | 15 | 24 | 21 | 30 | 120 | 12.1 | 30.0 | 10 | 20 | 2 |
| Vera re | 46 | 14 | 3 | 26 | 4 | 27 | 120 | 67.1 | 78.0 | 26 | 43 | 3 |
| Vera li | 15 | 1 | 17 | 3 | 38 | 46 | 120 | 84.2 | 88.0 | 26 | 45 | 4 |
| Fred re | 3 | 27 | 36 | 2 | 0 | 52 | 120 | 117.1 | 110.0 | 32 | 52 | 5 |
| Fred li | 21 | 18 | 17 | 33 | 27 | 4 | 120 | 24.4 | 42.0 | 16 | 29 | 2 |
| Ronja re | 28 | 21 | 19 | 10 | 15 | 27 | 120 | 12.0 | 32.0 | 10 | 18 | 1 |
| Ronja li | 29 | 18 | 14 | 25 | 23 | 11 | 120 | 11.8 | 34.0 | 9 | 18 | 0 |
| Julia re | 28 | 32 | 28 | 23 | 3 | 6 | 120 | 38.3 | 62.0 | 17 | 29 | 3 |
| Julia li | 18 | 29 | 25 | 38 | 9 | 1 | 120 | 45.8 | 64.0 | 19 | 37 | 3 |
| Laura re | 6 | 27 | 29 | 20 | 11 | 27 | 120 | 22.8 | 46.0 | 14 | 23 | 1 |
| Laura li | 24 | 16 | 21 | 16 | 12 | 31 | 120 | 11.7 | 32.0 | 11 | 19 | 1 |
| | 17% | 17% | 16% | 18% | 14% | 19% | | | | | | |

**Abb. 10b**: Nur bei Svenjas Bleistiftwürfel „rechtsherum" gibt keine Testgröße Anlass zum Zweifel an der Gleichwahrscheinlichkeit.

| | kritische Grenze | | | | | | | 11 | 30 | 11 | 18 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Σ (H-20)²/20 | Σ\|H-20\| | MAX\|H-20\| | MAX(H)-MIN(H) | #≤10+#≥30 |
| Nicola | 25 | 16 | 19 | 20 | 23 | 17 | | t | s | g | u | a |
| Wlad | 22 | 19 | 12 | 21 | 20 | 26 | 120 | 5.3 | 18.0 | 8 | 14 | 0 |
| Nilita | 19 | 19 | 21 | 21 | 19 | 21 | 120 | 0.3 | 6.0 | 1 | 2 | 0 |
| Flo | 20 | 17 | 18 | 24 | 24 | 17 | 120 | 2.7 | 16.0 | 4 | 7 | 0 |
| Thea | 20 | 22 | 23 | 16 | 17 | 22 | 120 | 2.1 | 14.0 | 4 | 7 | 0 |
| Nele | 24 | 24 | 17 | 17 | 16 | 22 | 120 | 3.5 | 20.0 | 4 | 8 | 0 |
| Mirko | 20 | 22 | 18 | 11 | 24 | 25 | 120 | 6.5 | 22.0 | 9 | 14 | 0 |
| Simon | 18 | 23 | 23 | 17 | 26 | 13 | 120 | 5.8 | 24.0 | 7 | 13 | 0 |
| Darius | 13 | 19 | 25 | 16 | 22 | 25 | 120 | 6.0 | 24.0 | 7 | 12 | 0 |
| Omar | 18 | 20 | 20 | 21 | 22 | 19 | 120 | 0.5 | 6.0 | 2 | 4 | 0 |
| Inga | 17 | 24 | 17 | 24 | 15 | 23 | 120 | 4.2 | 22.0 | 5 | 9 | 0 |
| Claudia | 17 | 24 | 25 | 12 | 25 | 17 | 120 | 7.4 | 28.0 | 8 | 13 | 0 |
| Ulla | 21 | 18 | 21 | 16 | 23 | 21 | 120 | 1.6 | 12.0 | 4 | 7 | 0 |
| Ayse | 19 | 17 | 9 | 21 | 24 | 30 | 120 | 12.4 | 30.0 | 11 | 21 | 2 |
| Max | 16 | 24 | 19 | 20 | 13 | 28 | 120 | 7.3 | 24.0 | 8 | 15 | 0 |
| Selma | 16 | 20 | 19 | 20 | 21 | 24 | 120 | 1.7 | 10.0 | 4 | 8 | 0 |
| | 16% | 17% | 16% | 15% | 17% | 18% | | | | | | |

**Abb. 11b**: Nur bei einer 120er-Stichprobe sind die Testwerte auf Gleichverteilung erhöht.

Ein Blick auf die **Abbildungen 9**, **10** und **11**, bei denen kritische Testgrößen unterlegt sind, zeigt Folgendes:

a) Wenn eine Testgröße erhöhte Werte signalisiert, sind i. d. R auch die anderen Testwerte erhöht. Wenn bei einer Testgröße Zweifel zu keimen beginnen, dann auch bei den anderen.

b) Die Testgrößen $t$ und $s$ reagieren i. A. etwas „sensibler“ auf Abweichungen von der Gleichverteilung als $g$, $u$ und $a$, wobei es auch Ausnahmen gibt, (vgl. **Abb. 11b**, Ayse).
In Bezug auf die Anwendungsbeispiele kann man festhalten:

c) Bei Stichproben mit 120 Gummibären signalisieren die Testgrößen nur selten Zweifel an der Gleichverteilung der Farben. Die Gleichverteilung ist ein für den Hausgebrauch gutes Modell.

d) Es gibt nur sehr wenige Sechskantbleistifte, für die die Gleichverteilung der sechs Seiten beim Rollen ein brauchbares Modell darstellt. Und wenn sie für eine Rollrichtung brauchbar ist, dann nicht notwendig auch für die andere Richtung.
Die letzte Zeile in **Abb. 10b** bestätigt aber die Brauchbarkeit des LAPLACE-Modells, wenn man die Ergebnisse über eine Vielzahl verschiedener Stifte mittelt. Alles andere würde dem gesunden Menschenverstand widersprechen.

e) Bei einer Neigung des Glücksrades um 10 % signalisieren die Testgrößen bei 120 Schnipsern verlässlich Zweifel an der Brauchbarkeit des LAPLACE-Modells.

f) Wenn man die kumulierten Daten analysiert, schält sich das folgende Modell als vertrauenswürdig heraus: Den Feldern 1 (oben) und 6 (unten) ordnet man $p(1) = p(6) = 1/6$ zu, den „bergauf“ – Feldern, das sind bei Drehung im Uhrzeigersinn die Felder 2 und 4, ordnet man $p(2) = p(4) = 1/6 + \alpha$ und den bergab Feldern $p(3) = p(5) = 1/6 - \alpha$ zu, wobei $\alpha = 5$ % eine gute Wahl ist. Tatsächlich wächst α proportional zur Steigung der schiefen Ebene [2], vgl. RIEMER [2017].

g) Erste Konsequenz: Wenn man die Drehrichtung zufällig oder systematisch wechselt, dann erweist sich die Gleichverteilung der sechs Felder trotz der Neigung als ausgezeichnetes Modell. Das gilt bis zu einer Steigung von ca. 30 % – solange der Zeiger auch bergab aufgrund der Reibung überall anhalten kann und nicht durch den Einfluss der Schwerkraft selbstständig nach unten rutscht.

h) Zweite Konsequenz: Wenn man 1 und 6 zum Ereignis A = {1;6}, 2 und 3 zum Ereignis B = {2;3} und 4 und 5 zum Ereignis C = {4;5} zusammenfasst, dann ist die Gleichverteilung für A, B, C ein ausgezeichnetes Modell. Ebenso $p(\text{oben}) = p(\{1;2;3\}) = p(\text{unten}) = p(\{3;4;5\}) = 1/2$.

## 5 Gleichnis

Die beim Entdecken und Erproben von Sortier- und Testgrößen gesammelten Erfahrungen lassen sich in einem einprägsamen Gleichnis festhalten, das die Philosophie des Hypothesentestens wie folgt auf den Punkt bringt:

In der Welt gibt es unzählige Hypothesen, die sich darum bewerben, als gute Modelle der Wirklichkeit zu gelten. Eine Testgröße wirkt wie ein Sieb, in dem schlechte Hypothesen (wie Kieselsteine) hängen bleiben. Sie werden wegen Überschreiten von Benchmarks bezweifelt. Nur die besseren bestehen den Test, sie werden vom Sieb durchgelassen. Je kleiner man die Löcher durch Verkleinern der Benchmarks macht, desto häufiger bleiben aber auch eigentlich brauchbare Hypothesen im Sieb hängen. Wenn man die Löcher vergrößert (die Benchmarks heraufsetzt), rutschen auch viele weniger gute Hypothesen hindurch. Und zu allem Überfluss gibt es verschiedene Testgrößen ($g$, $s$, $q$, $t$, $u$, $a$), die verschiedenen Sieben entsprechen. Oft geben einige Testgrößen Anlass für Zweifel, andere nicht. Dann sollte man sich „im Zweifel für den Zweifel“ entscheiden … und sich nach besseren Modellen umsehen.

## 6 Rückblick

Aus dem Alltag sind uns „Vertrauen und Zweifel“ geläufige Begriffe. Im Kontext des Würfelns und des Vertrauens in die Fairness von Würfeln gelingt es Schülerinnen und Schülern überraschend leicht, ihre Zweifel intuitiv zu ordnen, durch selbst erfundene Sortiergrößen, *die durch das Festlegen von kritischen Grenzen zu Testgrößen werden*, numerisch zu erfassen, zu vergleichen – und dadurch „Bauchgefühl“ im besten Sinne des Wortes zu „modellieren“.

Dabei sind Würfel mit ihren sechs möglichen Ergebnissen ungleich motivierender als Münzen mit nur zwei Ergebnissen, weil Schüler nicht nur eine, sondern viele sinnvolle Sortiergrößen finden – und erleben, was es bedeutet, wenn einige, möglicherweise aber nicht alle, Testgrößen Zweifel an der Gleichverteilung signalisieren. Auf diese Weise wachsen im Sinne von [SACHS 2009] Grundvorstellungen zur inhaltlichen Bedeutung/Aussagekraft der Testgrößen in der beurteilenden Statistik.

Die Rahmung durch die Deutung von Wahrscheinlichkeiten als Modellen der Wirklichkeit, die unter Berücksichtigung von Testgrößen nicht akzeptiert/verworfen, sondern modifiziert/verbessert werden, blockiert gängige Fehlvorstellungen, die sich beim rezepthaften Anwenden einseitiger Signifikanztests mehr oder weniger zwangsläufig zu entwickeln scheinen.

Nach Paradigma „keep it small and simple“ (KISS) wurde hier bewusst nur die Gleichverteilung als Modell untersucht – und auch nur für eine feste Stichprobengröße $n = 120$[1].

Wer mehr Zeit investieren möchte, kann mithilfe von Computersimulationen den Einfluss der Erhöhung von Stichprobengrößen untersuchen und bei Schülern die Erkenntnis wachsen lassen, dass man bei sehr hohen Stichprobengrößen praktisch jedes Modell bezweifeln (und im Sinne von Signifikanztests auch verwerfen) muss, obwohl es für den Hausgebrauch nützlich und mehr als ausreichend sein kann. „Tests sind entgegen einer ebenfalls verbreiteten Fehlvorstellung eben nicht für zu hohe Stichproben gemacht.“ Sie bei sehr großen Stichprobenumfängen dennoch zu verwenden, läuft auf das Beanstanden präzise gesägter Holzbretter hinaus, deren Länge man im Baumarkt mithilfe eines – der Sache völlig unangemessenen – Elektronenmikroskops kontrolliert. Und das wäre auch eine Antwort auf die schwierige Frage 2.2 b).

### Literatur


[1] AMRHEIN, V. GREENLAND, S. MCSHANE, B. [2019]: Retire statistical significance. Nature 567, S. 305 – 307. 2019
[2] BRANDT, D. HOCHE, D., RIEMER, W. WOLLMANN, W. [2018]: Lambacher-Schweizer, Themenband Stochastik – Hessen. Stuttgart 2018. Klett Verlag.
[3] BUTH, M. [2002]: Anmerkungen zum Testen von Hypothesen. Stochastik in der Schule 2002/2, S. 27 – 29.
[4] DIEPGEN, R. [2002]: Wie man das Testen von Hypothesen lieber doch nicht einführen sollte. Stochastik in der Schule 2002/3, S. 34 – 38.
[5] EICHLER, A, VOGEL, M. [2012]: Leitidee Daten und Zufall. Wiesbaden 2009, 1. Aufl. Teubner Verlag.
[6] KRAUSS, S., WASSNER, C. [2002]: Wie man das Testen von Hypothesen einführen sollte. Stochastik in der Schule 2001/1 S. 29 – 34.
[7] LERGENMÜLLER, A., SCHMIDT, G., KRÜGER, K. (Hrsg.) [2012]: Neue Wege Stochastik. Braunschweig 2012. Schroedel Verlag
[8] MOTZER, R. [2017]: Baumdiagramm beim Signifikanztest. Stochastik in der Schule 2017/1 S. 29 – 31.
[9] HENZE, N. [2017]: Stochastik für Einsteiger (2017). Berlin, Springer.
[10] MOßBURGER, M. [2014]: Unklare Begriffe und Wunschdenken bei Signifikanztests. Stochastik in der Schule 2014/1 S. 2 – 8.
[11] RIEMER, W. [1986]: Eine neue Sicht der BAYESschen Regel. Stochastik in der Schule 1986/3 S. 4 – 13.
[12] RIEMER, W. [2016]: Mit Quadern würfeln. Mathematik 5 – 10 Nr. 2016/36 S. 23 – 25. Friedrich Verlag.
[13] RIEMER, W. [2017]: Das Glücksrad auf der schiefen Ebene. In MAITZEN, S., WARMELING, A. (Hrsg.) Mathe aus dem Leben – für das Leben. MUED Festschrift 2017 ISBN 978-3-930197-90-3.
[14] RIEMER, W. [2020]: Signifikanztests und das Risiko falscher Entscheidungen: Gewinn besiegt Wahrheit. Mathematiklehren 2020.
[15] RIEMER, W. [2019]: Konfidenzintervalle. Handreichung im Thüringer Schulportal. https://www.schulportal-thueringen.de/web/guest/media/detail?tspi=3498.
[16] SACHS, L. [1999]: Angewandte Statistik. Berlin 1999, 9. Aufl. Springer Verlag.
[17] STOYAN, D. [2011]: Statistische Tests in Gymnasiallehrbüchern. Stochastik in der Schule 2011/1, S. 28 – 32.

## Anmerkungen

[1] Zur Prüfung beliebiger Hypothesen ($p_1$, $p_2$, …, $p_6$) nutzt man standardmäßig die $\chi^2$-Testgröße $t = \frac{(n_1 - n\cdot p_1)^2}{n\cdot p_1} + \frac{(n_2 - n\cdot p_2)^2}{n\cdot p_2} + \ldots + \frac{(n_6 - n\cdot p_6)^2}{n\cdot p_6}$ . Sie ist „genial“, weil die kritischen Grenzen (asymptotisch) weder von den Wahrscheinlichkeiten noch von der Stichprobengröße abhängen. Das kann man auf der Schule zwar nicht begründen, aber dadurch stützen, dass man zeigt: Der Erwartungswert von $t$ ist als Summe der Varianzen von sechs (voneinander abhängigen) binomialverteilten Zufallsvariablen konstant

$$E(t) = \frac{n\cdot p_1(1-p_1)}{n\cdot p_1} + \ldots + \frac{n\cdot p_6(1-p_6)}{n\cdot p_6} = 5\,,$$

also unabhängig von den Wahrscheinlichkeiten und dem Stichprobenumfang.

| | n | rot | gelb | grün | Nebenr. | | d |
|---|---|---|---|---|---|---|---|
| Cristian | 62 | 30 | 17 | 15 | 2 | 6 | 2 | 10 |
| Johann | 48 | 27 | 17 | 4 | 6 | 20 | 32 | 58 |
| | | \|2 · 27 − 48\| | | \|4 · 17 − 48\| | | \|4 · 4 − 48\| | | Σ |

$$d = |2n_1 - n| + |4\cdot n_2 - n| + |4\cdot n_3 - n|$$

$$d = \left|\frac{1}{p_1}n_1 - n\right| + \ldots + \left|\frac{1}{p_3}n_3 - n\right|$$

$$d = \frac{|n_1 - np_1|}{p_1} + \ldots + \frac{|n_3 - np_3|}{p_3} \quad \text{durch n teilen}$$

$$d' = \frac{|n_1 - np_1|}{np_1} + \ldots + \frac{|n_3 - np_3|}{np_3}$$

$$\chi^2 = \frac{|n_1 - np_1|^2}{np_1} + \ldots + \frac{|n_3 - np_3|^2}{np_3}$$

5 ist die Anzahl der Freiheitsgrade der zugehörigen $\chi^2$-Verteilung.

Als es in Klasse 6 vor (!) der Bruchrechnung einmal um die Frage ging, welche von mehreren Löffelstichproben aus einem Topf mit 2000 roten, 1000 gelben und 1000 grünen Perlen am besten zum Inhalt passt (Frage nach Repräsentativität), erfanden Johann und Christian ein ganzzahliges Abweichungsmaß $d$, das nach einer Termumformung die Nenner in der $\chi^2$-Formel auch ohne Rückgriff auf Varianzen plausibel macht. Statt durch Quadrate haben die beiden natürlich betragliche Differenzen verwendet. Die Parallelität zu $\chi^2$ ergibt sich dann in der SI durch eine Termumformung:

[2] Die Wahrscheinlichkeit, dass der Zeiger an der Stelle $0 \le \varphi < 2\pi$ stehen bleibt, wird durch folgende sinusförmige Wahrscheinlichkeitsdichte beschrieben:

$$f(\varphi) = \frac{1}{2\pi}\cdot\left(1 + k\cdot\sin(\varphi)\right) \text{ mit } k = \frac{\tan(\alpha)}{\rho} < 1\ .$$

Dabei muss die Steigung $\tan(\alpha)$ der schiefen Ebene kleiner sein als der Reibungskoeffizient $\rho$, der angibt, welcher Anteil des Gewichts als Bremskraft auf den Zeiger wirkt. Genau dann kann der Zeiger auch „bergab“ überall anhalten – vgl. RIEMER [2017].

Wegen $\rho \approx 0{,}3$ trägt die Theorie bis zu einer Steigung von 30 %.

[3] Tatsächlich ist die KOLMOGOROFFsche (axiomatische) Wahrscheinlichkeitsauffassung lernpsychologisch viel eingängiger als die frequentistische nach v. MISES. Man muss sie nur an konkrete Situationen und an subjektivistische Vorerfahrungen koppeln. Das ist im Rahmen der der Strukturmathematik-Welle leider nie geschehen. Der Fokus lag seinerzeit auf der „Ereignisalgebra“ als Anwendungsfeld der Mengenlehre, also auf Strukturen, weniger auf Inhalten.