# Überzeugen statt Beweisen – der zentrale Grenzverteilungssatz im Gymnasialunterricht

TIBOR NEMETZ, JUDITH SIMON, NORBERT KUSOLITSCH, BUDAPEST–WIEN

***Zusammenfassung:*** *Erfahrungsgemäß ist die für Wahrscheinlichkeitstheorie vorgesehene Zeit in der Lehrerausbildung, sowohl in Ungarn als auch in Österreich, zu knapp bemessen. Dies führt dazu, dass ein exakter Beweis des zentralen Grenzverteilungssatzes nur auf Kosten anderer Themen geführt werden könnte. Erschwerend kommt dazu, dass den Lehramtsstudenten oft ein ausreichender mathematischer Hintergrund fehlt.*

*Wir schlagen deshalb einen anderen Weg vor, nämlich das Verhalten des Zufalls an Beispielen zu demonstrieren, die die Studenten selbst üben können. Dabei nutzen wir die durch das Internet gegebenen Möglichkeiten über beliebige Datenmengen verfügen zu können.*

## 1 Einführung

Unter Stochastiklehrern ist es eine weit verbreitete Ansicht, dass eine Vorlesung über Wahrscheinlichkeitstheorie keinen Sinn macht, wenn

- das Gesetz der großen Zahlen nicht bewiesen wird,
- den Studenten die Grundidee statistischer Entscheidungen am Ende des Kurses nicht klar geworden ist,
- der zentrale Grenzverteilungssatz nicht erklärt wird.

Unserer Meinung nach genügt für den ersten Punkt nach einer gewissen Vorbereitung etwa eine Stunde, und dieses Thema wird auch in vielen Textbüchern in zufriedenstellender Weise behandelt. Mit dem zweiten Punkt wollen wir uns in einer anderen Arbeit beschäftigen.

Hier wollen wir zeigen, dass man den dritten Punkt in der Schule erfolgreich behandeln kann. In einem Übersichtsartikel von Le Cam (Le Cam, L. (1986)) ist ein sehr eleganter von Lindeberg stammender Beweis des zentralen Grenzverteilungssatzes dargestellt, der ohne charakteristische Funktionen auskommt. Trotzdem, glauben wir, kann man keinen exakten mathematischen Beweis in voller Allgemeingültigkeit führen. Weder die Zeit noch der mathematische Hintergrund reichen dafür aus. Stattdessen empfehlen wir die Verwendung von wirklich überzeugenden Beispielen. Dazu benötigt man eine große Menge statistischer Daten, mit denen man noch vor einigen Jahren nicht in einem Klassenzimmer hätte arbeiten können. Es hätte sowohl an der Verfügbarkeit entsprechend großen Datenmaterials als auch an den technischen Hilfsmitteln zu deren Analyse gefehlt. Aus dem Internet kann man sich jede Menge an Daten holen, und mittlerweile besitzen auch billige PCs hinreichend Rechen- und Speicherleistung, um große Datenmengen zu bearbeiten. Darüber hinaus gibt es viele leistungsfähige statistische Programmpakete, von denen einige auch frei verfügbar sind.

Wir wollen hier zeigen, wie man diese drei Faktoren nutzen kann, um den Schülern die Aussage des zentralen Grenzverteilungssatzes in überzeugender Weise zu demonstrieren.

Wir wiederholen den zentralen Grenzverteilungssatz in der Form, wie er für Lehramtsstudenten üblicherweise formuliert wird (siehe z.B. Renyi, A. (1970)):

*Sei eine unabhängig, identisch verteilte Folge von Zufallsvariablen gegeben und sei $T_n$ die Teilsumme der ersten n Zufallsvariablen, dann hat die zentrierte Version der $T_n$ eine Grenzverteilung, nämlich die Normalverteilung.*

Unser Ziel ist es nun, diese Aussage in einer allgemein verständlichen Weise zu präsentieren. Dies soll zunächst im Prinzip kurz skizziert werden und wird dann im nächsten Abschnitt im Detail behandelt:

- Beziehe digitale Dokumente aus dem Internet.
- Wandle sie in eine Zahlenfolge um und durchmische die Zahlenfolge.
- Wähle eine Blocklänge $b$, nimm aufeinanderfolgende Blöcke und berechne die zentrierten Teilsummen $T_n$.
- Berechne die empirische Verteilung dieser Folge von zentrierten Teilsummen $T_n$.
- Stelle diese empirischen Verteilungen grafisch dar und vergleiche sie für verschiedene Datenmengen.

## 2 Der Gebrauch des Internets zur Datenbeschaffung

Der Einfachheit halber beschränken wir uns auf geschriebene Texte aus dem Internet. Es gibt jeden Tag eine Unmenge an neuem Datenmaterial, z.B. aus den Internetversionen der Zeitungen. Am einfachsten sind klarerweise Textfiles zu behandeln, aber die meisten Browser bieten auch die Möglichkeit, HTML-Seiten als Text-Files zu speichern. Unabhängig davon, welches Datenformat man verwendet, sollte man versuchen, das File in ein Textformat umzuwandeln und alle Formatierungsbefehle zu löschen, ebenso wie eventuell eingefügtes Bildmaterial.

Um die Schüler davon zu überzeugen, dass nicht manipuliert wird, verwendet man immer wieder neues Datenmaterial. Dieser Effekt kann noch dadurch verstärkt werden, dass man den Schülern innerhalb gewisser technischer Grenzen freie Hand bei der Umwandlung der Textfolgen in Zahlenfolgen lässt, bspw. kann man alle Buchstaben durch beliebige 3-stellige Kommazahlen aus (0, 1) kodieren, wobei dem Schüler die Wahl der einzelnen konkreten Zahlen überlassen bleibt. Genauso gut könnte man auch die 2-stelligen Zahlen 01,...,26 etwa für das lateinische Alphabet nehmen, aber aus technischen Gründen ist es sinnvoll, Zahlen mit einer fixen Anzahl von Stellen und aus einem beschränkten Bereich zu wählen.

Die Folge sollte danach aber etwas durchmischt werden, damit keine allzu großen Abhängigkeiten zwischen benachbarten Zeichen bestehen.

## 3 Statistische Behandlung der Daten

Im folgenden wollen wir annehmen, dass der Text aus Zeichen eines gegebenen Alphabets $a_1$, $a_2$, ..., besteht. Wir wählen 2 Parameter und zwar die Blocklänge $b$ und die Anzahl der Blöcke $n$ (dementsprechend brauchen wir einen Text von mindestens $b \cdot n$ Buchstaben).

Für die Summe der Zahlen im $i$-ten Block verwenden wir die Bezeichnung $T_i$, $i = 1,...,n$. Sodann berechnen wir Stichprobenmittel $\overline{T}_{n,b} := \sum_{i=1}^{n} \frac{T_i}{n}$ und Stichprobenvarianz $S^2_{n,b} := \frac{1}{n-1} \Sigma_{i=1}^{n} (T_i - \overline{T}_{n,b})^2$.

Danach bilden wir die Folge der zentrierten Summe $Z_i := \frac{T_i - T_{n,b}}{S_{n,b}}$. Wir stellen die empirische Verteilung der zentrierten Summen durch ein Histogramm dar, wobei wir Werte außerhalb des Intervalls [–4,4] zu einer Klasse zusammenfassen und innerhalb dieses Intervalls in gleich breite Klassen, etwa der Breite 0,5 einteilen. Als Faustregel wird man daher die Anzahl der Blöcke etwa mit $n \approx \frac{5}{\Phi(3) - \Phi(2,5))} \approx 5/0.004859767 \approx 1024$ festlegen. Es spielt jedoch von der Rechenzeit und vom Aufwand her keine Rolle, $n$ bspw. um den Faktor 10 zu vervielfachen.

Üblicherweise nehmen einem die modernen statistischen Programme die Arbeit der Klasseneinteilung ab, doch ist es normalerweise auch möglich, die Einteilung von Hand aus vorzunehmen.

Es wird empfohlen, das oben beschriebene Verfahren für verschiedene $b$, bspw. $b$ = 10,20,... durchzuführen, um den Einfluss der Blocklänge auf die Grafiken betrachten zu können. Dies wirkt besonders einprägsam, wenn man auch noch die Glockenkurve in die Grafik einzeichnet.

Sehr lehrreich ist es auch, Grafiken miteinander zu vergleichen, die aus verschiedenen Codierungen des Textes hervorgegangen sind.

In den folgenden Grafiken sind auch die Histogramme der ursprünglichen Daten dargestellt, um zu zeigen, dass die Verteilung der zentrierten Summen davon praktisch unabhängig ist.

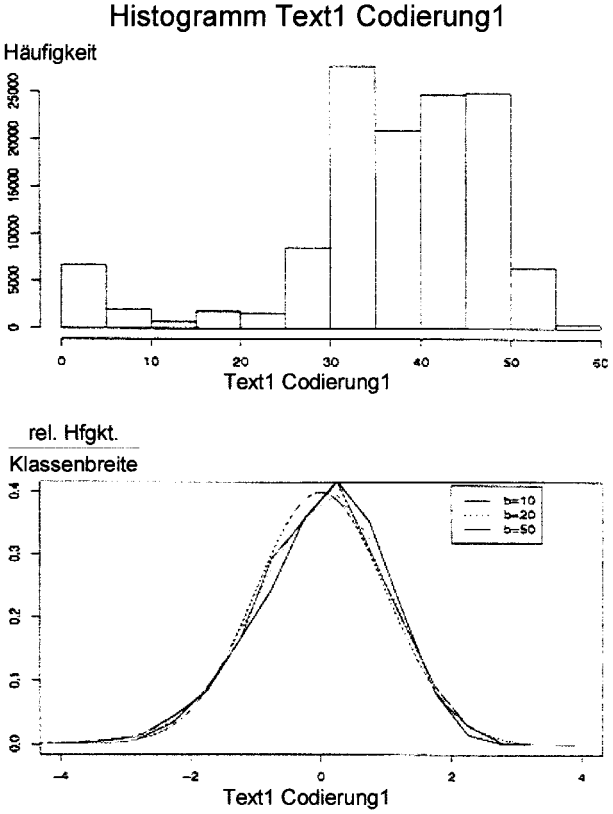



Abb. 1: Text1, Codierung1

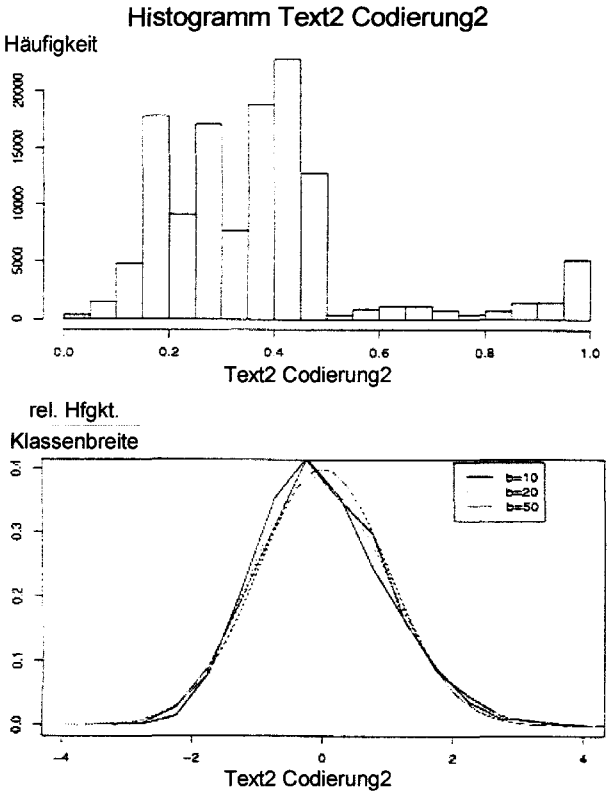



Abb. 2: Text2, Codierung2

# 4 Die Binomialverteilung

Auch für den Fall der Binomialverteilung ist die Verwendung des PCs sehr nützlich, denn selbst dafür übersteigt ein formaler Beweis das Gymnasialniveau, obwohl man für den symmetrischen Fall die Größenordnung der Abweichungen sehr elegant herleiten kann (siehe Nemetz–Kusolitsch (1999)). Ähnlich elementare Herleitungen für den unsymmetrischen Fall kennen wir nicht.

Bekanntlich ist eine Zufallsvariable $X$ binomialverteilt mit den Parametern $n$ und $p$, wenn gilt:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,\ldots,n$$

Die durch die Rechengenauigkeit gesetzten Grenzen machen eine direkte Berechnung der obigen Wahrscheinlichkeiten schon bei relativ kleinem $n$ unmöglich. Wir wollen hier eine Methode vorstellen, wie man die Wahrscheinlichkeiten sukzessive mit einfacher Genauigkeit auch noch für große $n$ ($n \leq 1024$) bestimmen kann. Wenn man diese Wahrscheinlichkeiten berechnet hat, so kann man die Verteilung der zentrierten Version einfach grafisch darstellen und mit der Normalverteilung vergleichen.

Wir beobachten ein alternativverteiltes Experiment mit den Ausgängen $A, \overline{A}$ mit den Wahrscheinlichkeiten $p := P(A)$, $q := 1-p$ $n$-mal. Sei $X$ die Häufigkeit von $A$ in $n$ Beobachtungen. Es ist bekannt, dass $X$ binomialverteilt mit $n$, $p$ ist. Sind $X$ und $Y$ unabhängig und binomialverteilt mit den Parametern $n$, $p$ bzw. $m$, $p$ so ist $Z = X + Y$ binomialverteilt mit den Parametern $n+m, p$. Dementsprechend schlagen wir folgenden Algorithmus vor:

**1. Schritt** (die Wahrscheinlichk. für Potenzen von 2)

**Initialisierung**

$P(i) := Q(i) := R(i) = 0,\ i = 0,\ldots,n;$
$P(0) := Q(0) := B(0,0) := q;$
$P(1) := Q(1) := B(0,1) := p;$

**Schleife**

Für $k = 1,\ldots,m = \lfloor \log_2 n \rfloor$ :
  Für $i = 0,\ldots,2^k$:
    $R(i) := \Sigma_{j=1}^{i} P(i)\cdot Q(i-j)$ ;
    $B(k,i) := P(i) := Q(i) := R(i)$
  Ende Schleife $i$.
Ende Schleife $k$

**Bemerkung** Damit erhält die $k$-te Zeile $B(k,i)$, $i = 0,\ldots,2^k$ der Matrix $B$ die Binomialverteilung mit den Parametern $2^k, p$.

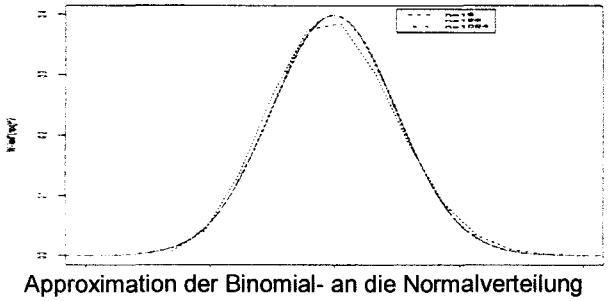
**2. Schritt** Bestimme die Binomialdarstellung von $n$, also $n = b_0 + 2b_1 + \ldots\ 2^m b_m$ und speichere die Indizes $k$ mit $b_k = 1$ in ein Feld $i(0),\ldots,i(l)$, $l \leq m$.

**3. Schritt** (Zusammensetzung der Binomialverteilung mit $n,p$)

$R(j) := B(i(0),j);\ j = 0,\ldots,2^{i(0)}$
Für $k = 0,\ldots,l-1$:
  $P(j) := R(j)\ \ j = 0,\ldots,2^{i(k)};$
  $Q(j) := B(i(k+1),j)\ \ j = 0,\ldots,2^{i(k)};$
  Berechne das Faltprodukt von $P$ und $Q$ und speichere es in $R$.
Ende Schleife $k$.

Am Ende der letzten Schleife enthält $R$ die gesuchte Binomialverteilung.

Auf Schulniveau ist es zumutbar, den Erwartungswert und die Varianz der Binomialverteilung zu berechnen. Damit ist es einfach, die zentrierten Werte zu bilden, über die man die mit $\sqrt{np(1-p)}$ multiplizierten Wahrscheinlichkeiten aufträgt und verbindet. So erhält man ein „*Wahrscheinlichkeitspolygon*“ der Binomialverteilung, das man wieder mit der Dichte der Normalverteilung vergleicht.



Approximation der Binomial- an die Normalverteilung

Abb. 3. Binomialverteilungen

Auch hier wird dem Lehrer empfohlen, die Grafiken für verschiedene $n$ und $p$ darzustellen.

## 5 Schlussbemerkungen

Welche Zahlenmenge benutzt wird, ist unwesentlich, da die Standardisierung immer auf denselben Trägerbereich führt.

Es ist wohlbekannt, dass die Normalapproximation nicht gut ist, wenn $np$ ungefähr konstant ist. In diesem Fall ist es natürlich sinnvoller, durch die Poissonverteilung zu approximieren. So ist bspw. bei $p = 0{,}2$ ein Stichprobenumfang von $n = 50$ zu klein für eine gute Normalapproximation. In diesem Zusammenhang sei auf die in Schulbüchern zumeist zitierte Faustregel $npq > 9$ verwiesen.

Verzichtet man auf eine Durchmischung der Datenmenge und bildet statt dessen Blöcke aus bspw. jedem fünften Buchstaben, so wird sich an den Ergebnissen nichts Wesentliches ändern, und die Studenten können erahnen, dass der zentrale Grenzverteilungssatz auch unter viel allgemeineren Voraussetzungen gilt als die, die etwa in der in Abschnitt 1 wiedergegebenen Fassung angeführt sind. So wichtig diese allgemeineren Formen in den praktischen Anwendungen auch sind, so ausgeschlossen bleibt ihre mathematisch exakte Behandlung im Rahmen eines Lehramtsstudiums. Dementsprechend wichtig scheint es uns, die Studenten zumindest von der Richtigkeit und Tragweite der Aussagen des zentralen Grenzverteilungssatzes überzeugen zu können.

## Literatur


Le Cam, L. (1986) The Central Limit Theorem around 1935. In: Statistical Science Vol. 1, No. 1 (1986)

Nemetz, T. – Kusolitsch, N. (1999) Guide to the Empire of Random. Verlag TypoTex, Budapest

Rényi, A. (1970) Probability Theory, Verlag Akadémiai Kiadó, Budapest


## Autoren


Norbert Kusolitsch
Institut für Statistik
Technische Universität Wien
E-Mail: kusolitsch@ci.tuwien.ac.at

Tibor Nemetz
Rényi-Institut
Ungarische Akademie der Wissenschaften
Budapest
E-Mail: nemetz@renyi.hu


Das Internet ist auch eine Fundgrube für einschlägige Briefmarken, Geldscheine, Münzen, Medaillen, Token usw. Drei Beispiele:


Die 10-jährige Kanadierin Claudia gestaltete diese Münze mit geometrisch-charakteristischen Kleinkinder-Körpern.

(Canada 1999, 25 Cents, Silber 5,8391g, KM #350a)


Zum „International Literacy Year" erschien diese Münze mit fleißig lesenden und schreibenden Kindern/Heranwachsenden.

(Canada 1990, 100 Dollars, Gold 13,3375g, KM #171)

„Eclectic geometric design" weisen die Münzkataloge hier aus. Das Histogramm verdeutlicht die stete aufstrebende Entwicklung.

(Canada 1999, 25 Cents, Silber 5,8391g, KM #353a)