Zentrale Sätze der Wahrscheinlichkeitsrechnung und damit verbundene fundamentale Ideen

MANFRED BOROVCNIK, KLAGENFURT

Der zentrale Grenzwertsatz (ZGS) zeichnet die Normalverteilung aus, welche dadurch eine Schlüsselrolle in der Stochastik erhält. Der ZGS rechtfertigt die Normalverteilung als Approximation für Zufallsvariable, welche die Summe anderer Zufallsvariablen sind oder als solche Summen gedacht werden können. Das zentrale Experiment in diesem Aufsatz hat mit Textanalyse aus statistischer Sicht zu tun. Es ist motivierend für Lernende, dass man die Gestalt der Verteilung, die untersucht wird, vorhersagen kann. Die Überlegungen motivieren auch, wie die stetige Standardnormalverteilung als Grenzwert von diskreten Verteilungen auftauchen kann. Textanalyse liefert einen natürlichen Kontext, in dem man die Beziehungen zwischen Stichproben und Population ansprechen kann, welche den Kern der beurteilenden Statistik ausmachen. Die Ausführungen sind an Borovcnik (2015) angelehnt.

1. Einleitung

Wahrscheinlichkeit ist ein feinsinniges Konzept, das leicht zu Missverständnissen führt. Anders als in der Geometrie ist unsere Wahrnehmung davon nicht trainiert worden, weil Wahrscheinlichkeit gar keine physikalische Eigenheit ist. Dennoch wird der Begriff häufig mit relativen Häufigkeiten eines Ereignisses in einer Serie von Experimenten gleichgesetzt. Tatsächlich gibt es eine – komplizierte – Beziehung zwischen diesen Begriffen (wenn man nur Experimente unter denselben Bedingungen wiederholen könnte). Statistiker ziehen es daher vor, Wahrscheinlichkeit als eine Metapher anzusehen, um über zufällige Situationen zu sprechen; oder sie halten Wahrscheinlichkeit für ein virtuelles Konzept, geradeso wie das Internet oder Computerspiele eine virtuelle Welt darstellen.

Aus mathematischer Sicht regeln drei Gruppen zentraler Sätze (mit vielen Varianten), was Wahrscheinlichkeit ist und wie man dies interpretieren kann: Das Gesetz der großen Zahlen (GGZ), der zentrale Grenzwertsatz (ZGS) und das Bayes-Theorem (BT). Das GGZ stützt eine Interpretation von Wahrscheinlichkeit als relative Häufigkeiten; vereinfachend kann man sagen, dass die relativen Häufigkeiten gegen die (unbekannte) Wahrscheinlichkeit konvergieren. Der ZGS erklärt, warum die Variation einer Zufallsvariablen in vielen Fällen durch eine Normalverteilung approximiert werden kann. Der einfachste Fall ist in der Geschichte der Stochastik als Fehlerverteilungsgesetz berühmt geworden. Danach kann man sich in einem Gedankenexperiment den Messfehler als Summe von (nicht beobachtbaren) elementaren Fehlern denken, sodass der resultierende Fehler, der beobachtet wird, einer Normalverteilung folgen sollte. Das BT zeigt, wie man qualitative Information (eine subjektive Einschätzung) über eine Wahrscheinlichkeit mit den relativen Häufigkeiten eines Zufallsexperiments verbindet. Daraus kann man ableiten, dass die revidierte subjektive Einschätzung einerseits zu einem Punkt und andererseits zu den relativen Häufigkeiten konvergiert. (Borovcnik, 2013).

Die vereinfachende Aussage zum GGZ ist irreführend, aber sie hat einen wahren Kern. Wir könnten das Theorem mathematisch analysieren. Die didaktische Frage ist vielmehr, wie man Szenarios (mit begleitenden Bildern) einführt, mit denen man das Thema passend unterrichten und seine Bedeutung klarmachen kann, wobei man stabile Intuitionen aufbaut, welche einerseits mit dem mathematischen Hintergrund im Einklang stehen und andererseits falsche Vorstellungen revidieren lassen. Wie kann man auf intuitiver Ebene klären, unter welchen Bedingungen die relativen Häufigkeiten gegen die zugrundeliegende Wahrscheinlichkeit konvergieren? Die vereinfachenden Aussagen zum ZGS sind falsch, da die Summe der Elementarfehler gar nicht konvergieren kann, weil sie mit zunehmender Zahl der Elementarfehler tendenziell immer größer wird – und das noch mit einer wachsenden Variabilität. Die Herausforderung ist, Situationen gezielt zu untersuchen, Arten von Konvergenz und Divergenz zu thematisieren und zu klären, von welcher Art diese Konvergenz zur Normalverteilung ist.

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 48, 2015, S 14 - 29

Wir nutzen Simulationen von Zufallsexperimenten und didaktische Animationen von Binomialverteilungen, um die "Daten" aus verschiedensten Perspektiven zu analysieren und zulässige Ideen über den ZGS zu stützen, welche verstehen lassen, wie der Begriff von Wahrscheinlichkeit nützlich werden kann, wenn man Informationen aus Daten extrahieren und verallgemeinern will.

2. Analyse eines gewöhnlichen Textes

Textanalyse und Interpretation ist ein subtiles Gebiet der Linguistik. Wir werden dagegen Textanalyse in sehr verengter Sicht verstehen. Wir ordnen den Zeichen eines Textes numerische Codes zu und analysieren u.a. die Häufigkeit der Codes oder die Verteilung der Code-Summe in kleineren Ausschnitten des Textes.

Der Leser mag sich an Zeiten als Kind zurückerinnern, als jemand einen "magischen" Trick an uns ausprobiert hat: „Denk dir zwei Zahlen zwischen 1 und 10 aus". Dann folgten Anleitungen zum Rechnen wie, addiere die zwei Zahlen, nimm das Quadrat der Summe, multipliziere nun mit 9, nimm schließlich die Wurzel, ziehe das Dreifache der zweiten Zahl ab und dividiere durch die erste Zahl. „Du musst jetzt 3 als Ergebnis haben!" sagte uns diese Person. Wir waren höchst überrascht.

Wir werden ein Experiment vorstellen, das auf der „Analyse" von Text aufbaut. Statt an zwei Zahlen zu denken, lassen wir die Testperson einen beliebigen Text von bestimmter Mindestlänge wählen. Anstelle der arithmetischen Operationen lassen wir die Verteilung der Zahlen, welche Textblöcken zugeordnet werden, untersuchen. Wir können vorhersagen, dass die Testperson eine Verteilung bekommt, welche der Standardnormalverteilung ziemlich ähnlich sieht.

2.1 Das Experiment

Wir folgen einer Empfehlung von Nemetz, Simon und Kusolitsch (2002). Nimm einen längeren Text, nach freier Wahl. Entferne Leerzeichen, Sonderzeichen (Interpunktion etwa). Stelle sicher, dass der verbleibende Text mindestens 20.000 Zeichen hat. Ordne den Text in einer Spalte in einer Tabellenkalkulation an (da können wir mit einem Trick helfen). Ordne jedem Zeichen einen Code zu, der aus Zahlen zwischen 1 und 1000 besteht (nach freier Wahl). Trenne die Zeichen in Blöcke der Länge 20. Berechne die Summe im ersten 20er-Block, und wiederhole die Berechnung für alle Blöcke.

Bei 20.000 Zeichen erhalten wir 1.000 Blöcke mit je 20 Zahlen und entsprechend 1.000 Daten für die Blocksumme. Wir berechnen den Mittelwert und die Standardabweichung dieser Daten. Damit berechnen wir die standardisierten Blocksummen, indem wir von den Blocksummen den Mittelwert abziehen und das Ergebnis durch die Standardabweichung dividieren. Damit erhalten wir 1.000 standardisierte Blocksummen. Nun sage ich voraus, dass beinahe alle Werte innerhalb der Grenzen von -5 und 5 liegen werden und ein Histogramm der Daten ziemlich nahe an der Standardnormalverteilungskurve liegen wird.

Lade zwei Freunde zu diesem Experiment ein. Sie sollten ihre eigene Zuordnung von Zahlen zu den Zeichen verwenden. Ihr Histogramm wird am Ende eurem ziemlich ähnlich sein und ganz nahe bei der Standardnormalkurve liegen. Wiederholt das Experiment mit 40.000 Zeichen und bildet nun Blöcke der Länge 40. Jetzt wird das Histogramm noch näher bei der Normalverteilungskurve liegen als zuvor. Ihr könnt auch einen anderen Text hernehmen. Wenn ihr die Zeichen im Text durch Zufall umordnet (das kann man leicht mit Zufallszahlen), dann wird die Übereinstimmung mit der Standardnormalverteilungskurve noch besser werden. Wie kommt es, dass ich das Ergebnis vorhersagen konnte. Das ist kein Trick; wie im Zahlenspiel von vorhin kann man das erklären. Die Erklärungen gehen jedoch weit über einfache Gleichungen hinaus und haben mit dem ZGS zu tun. Wir zeigen zuerst die Entwicklung des Spiels mit einem speziellen Text.

2.2 Spezifische Schritte der Analyse des Textes

Wir benutzen einen kürzlich veröffentlichten Aufsatz zu Risiko und ordnen den Zeichen ihre ASCII-Codes zu. In Tab. 1 zeigen wir nur das Ergebnis des Codierens im ersten 20er-Block. Wir berechnen die Summe in diesem Block und erhalten $b_1 = 2026$.

Links in Tab. 2 sind die Blocksummen der ersten zwanzig 20er-Blöcke gezeigt, nur um ein Gefühl für die Variation zu vermitteln. Aus allen Daten für die Blocksumme berechnen wir Mittelwert und Standardabweichung und erhalten (von unseren Daten, die in einem File erhältlich sind) $\overline{b} = 2143.32$ und $s_b = 33.08$.

Die erste standardisierte Blocksumme erhält man durch $\frac{b_1 - \overline{b}}{s_b} = \frac{2026 - 2143.32}{33.08} = -3.5469$.

Tab. 1: Textcodierung und Einteilung in Blöcke

Zeichen	Code	Nr.	Block Nr.	Pos. im Block
R	82	1	1	1
i	105	2	1	2
s	115	3	1	3
k	107	4	1	4
a	97	5	1	5
n	110	6	1	6
d	100	7	1	7
D	68	8	1	8
e	101	9	1	9
c	99	10	1	10
i	105	11	1	11
s	115	12	1	12
i	105	13	1	13
o	111	14	1	14
n	110	15	1	15
M	77	16	1	16
a	97	17	1	17
k	107	18	1	18
i	105	19	1	19
n	110	20	1	20

Tab. 2: Auswertung der Daten in den Textblöcken

Block Nr.	Block Summe	Mittel, Stand.abw.	Standardisierte Summe	Klasse $(e_{i-1}, e_i]$	Mitte m_i	Häufigkeit abs. n_i	Häufigkeit rel. h_i	Dichte h_i /0.2	Dichte Std.norm.
1	2026		–3.5469						
2	2015		–3.8794	–4.8	–4.9	0	0.000	0.000	0.000
3	2052	2143.32	–2.7608	–4.6	–4.7	1	0.001	0.005	0.000
4	2077	33.08	–2.0050	–4.4	–4.5	1	0.001	0.005	0.000
5	2097		–1.4004	–4.2	–4.3	0	0.000	0.000	0.000
6	2100		–1.3097	–4.0	–4.1	1	0.001	0.005	0.000
7	2143		–0.0096	–3.8	–3.9	1	0.001	0.005	0.000
8	2177		1.0183	–3.6	–3.7	2	0.002	0.010	0.000
9	2167		0.7159	–3.4	–3.5	2	0.002	0.010	0.001
10	2134		–0.2817	–3.2	–3.3	1	0.001	0.005	0.002
11	2116		–0.8259	–3.0	–3.1	2	0.002	0.010	0.003
12	2155		0.3531	–2.8	–2.9	2	0.002	0.010	0.006
13	2182		1.1694	–2.6	–2.7	5	0.005	0.025	0.010
14	2206		1.8950	–2.4	–2.5	3	0.003	0.015	0.018
15	2123		–0.6143	–2.2	–2.3	6	0.006	0.030	0.028
16	2179		1.0787	–2.0	–2.1	13	0.013	0.065	0.044
17	2173		0.8973	–1.8	–1.9	15	0.015	0.075	0.066
18	2075		–2.0655	–1.6	–1.7	8	0.008	0.040	0.094
19	2203		1.8043	–1.4	–1.5	19	0.019	0.095	0.130
20	2141		–0.0701	–1.2	–1.3	20	0.020	0.100	0.171

Wir setzen in den anderen Blöcken fort und erhalten 1.000 standardisierte Summen. Die relativen Häufigkeiten der Klassen (–5, –4.8], (–4.8, –4.6], …, (4.8, 5] seien mit h_i notiert, die Mittelpunkte mit m_i; wir berechnen die Häufigkeitsdichte durch Division mit der Klassenbreite (0.2) und zeichnen ein Häufigkeitspolygon, das die Punkte (m_i, h_i/0.2) verbindet (rechte Seite, Tab. 2). Wir ziehen ein Häufigkeitspolygon einem Histogramm vor, da es eine klarere Interpretation als Funktion liefert und wir dieses Polygon mit der Dichte der Standardnormalverteilung vergleichen. Wir zeigen nur einen kleinen Ausschnitt aus der Häufigkeitstabelle.

Das Häufigkeitspolygon in Abb. 1 ist der Standardnormalkurve schon ziemlich ähnlich. Die Anpassung könnte jedoch, besonders in der Mitte, etwas verbessert werden. Das wird durch die Besonderheiten des Textes, der bestimmte Zeichenfolgen begünstigt, verursacht.

2.3 Die Anpassung an die Normalverteilung wird besser, wenn der Text zufälliger wird

Wir wiederholen die Analyse, wobei wir den Originaltext durch eine Zufallsfolge umordnen (siehe weiter unten). Das Häufigkeitspolygon (Abb. 2) zeigt eine deutliche Verbesserung der Anpassung.

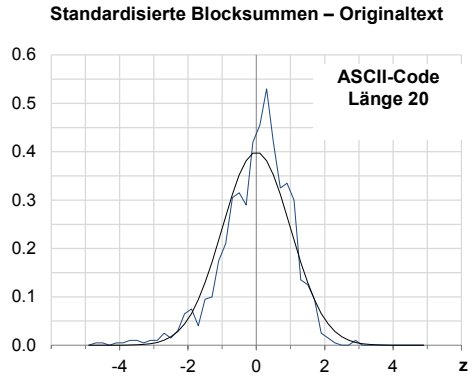

Abb. 1: Häufigkeitspolygon für standardisierte Blocksummen im Originaltext – ASCII-Codes für einzelne Zeichen.

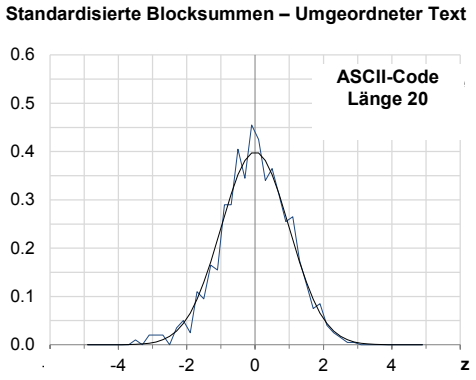

Abb. 2: Polygon der Häufigkeitsdichte für die standardisierte Blocksumme im zufällig umgeordneten Text (ASCII-Codes).

2.4 Auswirkung des verwendeten Codierungsschemas

Man könnte vermuten, dass die Zuordnung der Codes die gute Anpassung an die Standardnormalkurve „verursacht“ hat. Wenn man jedoch die Verteilung der zugeordneten Codes betrachtet, so gibt die kein klares Bild, sie sieht ziemlich erratisch aus (Abb. 3). Nichts erinnert an eine Normalverteilung. Es gibt einige Ausreißer im Bereich zwischen 45 und 90, die sich ungleich über ein breites Intervall verteilen. Es scheint nun umso erstaunlicher, dass die Normalverteilung so gut passt.

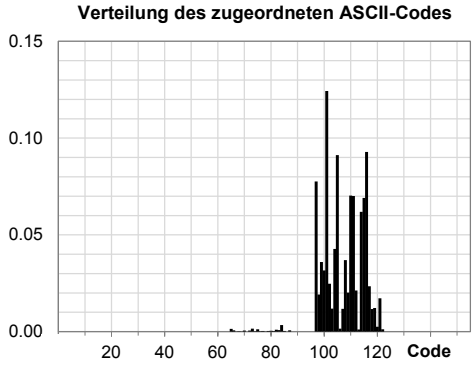

Abb 3: Verteilung der Codes für den ganzen Text von 20.000 Zeichen für den ASCII-Code.

Schauen wir uns die Ergebnisse der zwei anderen Personen an, welche die Zeichen des Textes anders codierten. Unterstellen wir, dass einer Folgenummern NR von 1 bis 55 verwendet hat, das ist relativ kompakt im Vergleich zum ASCII-Code, ohne Lücken. Für die andere Person nehmen wir an, dass 20.000 Zufallszahlen erzeugt und der Größe nach angeordnet wurden, sodass RD_i die i-t-kleinste Zufallszahl ist. Wenn ein Zeichen mit dem Folgenummern-Code NR mit i codiert wurde, dann ordnet der Zufallscode den ganzzahligen Teil von $i \cdot \mathrm{RD}_i \cdot 10$ zu. Diese Methode soll sicherstellen, dass der Code im Wesentlichen durch Zufall bestimmt ist.

Wir untersuchen die Verteilung der Blocksummen (mit Blocklänge 20) wie zuvor und erhalten Häufigkeitspolygone, die in etwa dieselbe Anpassung an die Standardnormalverteilungskurve zeigen. Wir zeigen nur die Ergebnisse für den zufällig umgeordneten Text (Abb. 4), weil die Anpassung durch

diese Umordnung deutlich verbessert wird. Für beide Codierungsschemata zeigt sich im ersten Intervall links von der Null eine leichte Überrepräsentanz, ansonsten eine sehr gute Anpassung.

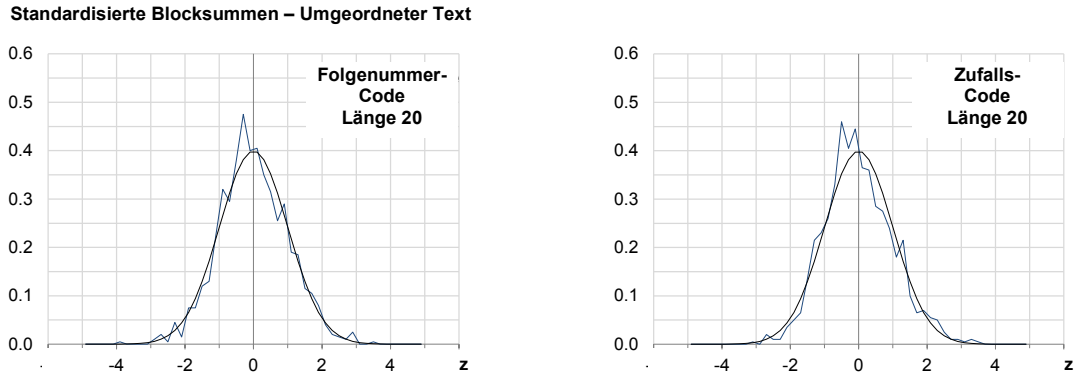

Abb. 4: Polygon der Häufigkeitsdichte für die standardisierte Blocksumme im zufällig umgeordneten Text. Links: Aufeinanderfolgende Zahlen als Codes. Rechts: Zufällig generierte Codezahlen.

Tab. 3 zeigt authentisch, wie den Zeichen des Textes Codes zugeordnet werden, auch wenn nur ein kleiner Teil gezeigt wird. Wie schon beim ASCII-Code können wir die Verteilung der einzelnen Codes im ganzen Text untersuchen. Bei der Folgenummer-Codierung ist die Verteilung sehr kompakt aber ziemlich ungleich; die zufällige Zuordnung hat eine viel größere Streuung (die erste Achse erstreckt sich von 0 bis 500 im Vergleich zu 0 bis 50 für die Folgenummern als Codes, Abb. 5). Dennoch ist das Ergebnis – die gute Anpassung der Normalkurve – gleich für beide Codierungsschemata.

Tab. 3: Teil der Codierungstabelle verschiedener Systeme, die in der Analyse angesprochen werden

Zeichen	ASCII	Nr.	Zufall	Zeichen	ASCII	Nr.	Zufall
A	65	1	0	C	67	5	10
a	97	2	1	c	99	6	12
B	66	3	3	D	68	7	14
b	98	4	4	d	100	8	17

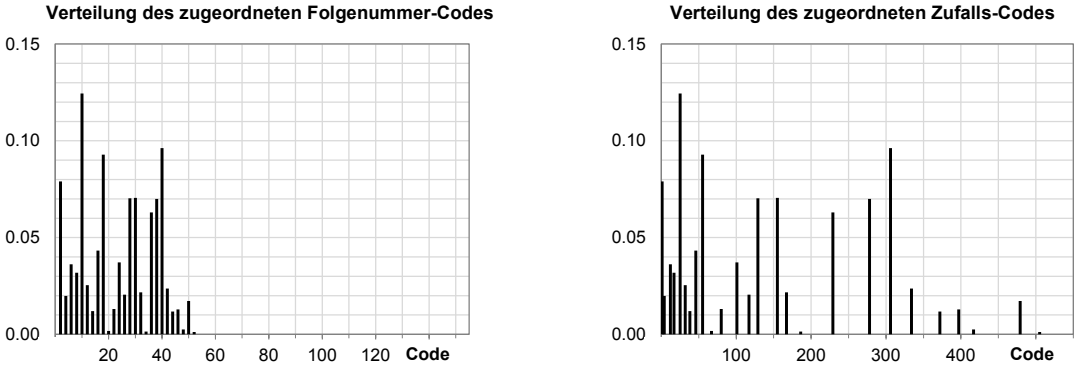

Abb 5: Verteilung der Codes im ganzen Text mit aufeinanderfolgenden Zahlen bzw. mit Zufallszahlen als Codes.

2.5 Auswirkung der Länge des Textes

Wir haben noch die Auswirkung der Länge des Textes auf die Verteilung der Blocksumme zu untersuchen. Obwohl es eine Verbesserung gibt (Abb. 6), hätten wir von der Theorie her mehr erwartet. Das rührt daher, dass der Text Eigenheiten hat, die nicht nur Abhängigkeiten zwischen den aufeinanderfolgenden Zeichen erzeugen (die aber durch zufälliges Umordnen ausgemerzt sein sollten), sondern auch die Zeichen in längeren Blöcken einschränken. Wenn der Text über Risiko ist, dann wird etwa

Risiko sehr oft genannt werden. Wir werden sehen, dass die Anpassung an die Standardnormalkurve beträchtlich verbessert wird durch Verdoppeln der Textlänge, wenn wir Text künstlich erzeugen.

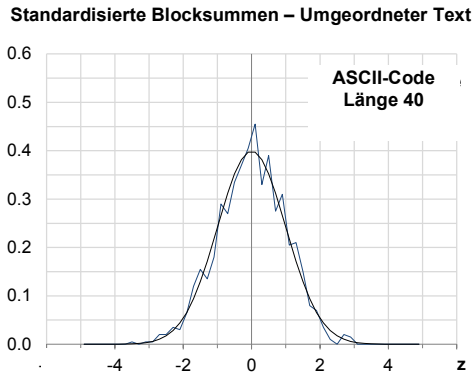

Abb. 6: Häufigkeitspolygon für die standardisierte Blocksumme im zufällig umgeordneten Text (ASCII-Codes) – Blocklänge 40.

3. Erzeugen von künstlichem Text mit nur zwei Zeichen

Anstelle natürlicher Texte werden wir jetzt Text erzeugen, sodass die probabilistischen Annahmen besser erfüllt werden. Wir verwenden nur 0 und 1 und erzeugen die Zeichen unabhängig voneinander; wir können uns ein Glücksrad mit zwei Sektoren vorstellen (siehe Abb. 7). Durch zwanzig Drehungen erzeugen wir einen Block der Länge 20. Wir wiederholen die Schritte 1.000 Mal, um die Textanalyse von vorhin zu imitieren.

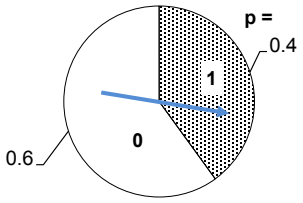

Abb. 7: Drehen des Glücksrads zwanzig Mal erzeugt einen Block (eine Stichprobe) der Länge 20.

3.1 Analyse von künstlichem Text

Wir erzeugen binären Text mit den Zeichen 0 und 1 zufällig; zuerst verwenden wir $p = 0.4$ für das Zeichen 1. Wir berechnen wie in Abschnitt 2 die Blocksummen und standardisieren die 1.000 Daten, die wir insgesamt erzeugen. Die Verteilung der standardisierten Werte wird wieder durch ein Häufigkeitspolygon dargestellt.

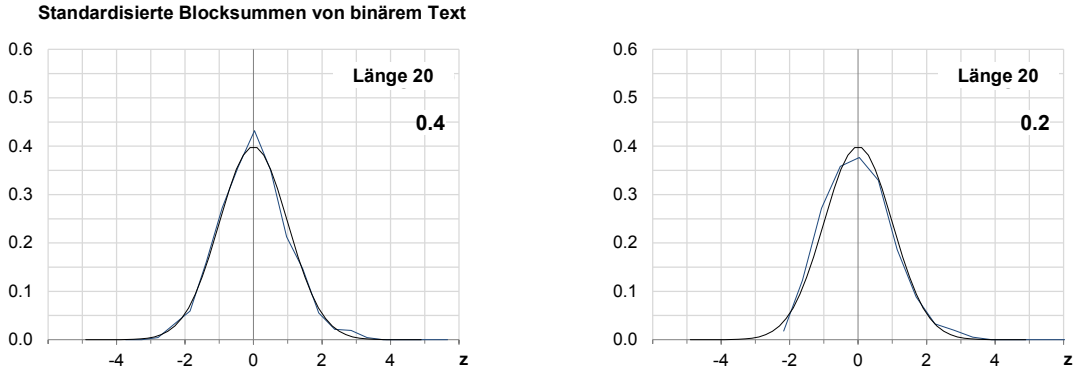

Abb. 8: Vergleich von Häufigkeitspolygon der standardisierten Blocksumme von binärem Text (0, 1) mit der Standardnormalkurve. Links: $p = 0.4$ für die 1. Rechts: $p = 0.2$ für die 1.

Es ist erstaunlich, wie gut die Anpassung ist (Abb. 8, links). Wenn wir einen Text mit einem kleineren Wert für p (0.2) erzeugen, so ist die Anpassung schlechter (Abb. 8, rechts), würde sich aber wieder

verbessern, wenn wir die Anzahl der Zeichen im Block vergrößern. Das Polygon zeigt eine systematische Ausbeulung nach links (linkssteil) verglichen mit der Standardnormalkurve. Wir replizieren die Texterzeugung durch Simulation und unterteilen 40.000 Zeichen nun in Blöcke der Länge 40. Um auch die Unschärfe der Simulation zu demonstrieren, zeigen wir zwei Häufigkeitspolygone mit $p = 0.4$ und mit $p = 0.2$. (Abb. 9).

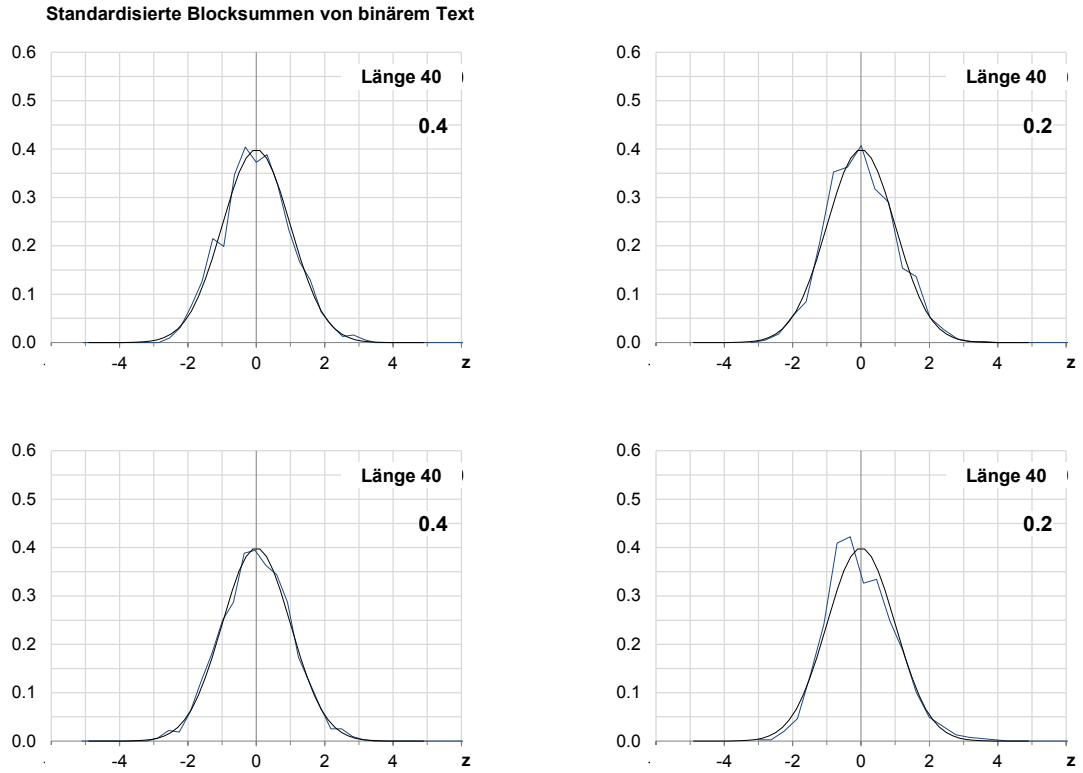

Abb. 9: Zwei Replikationen von binärem Text – Verteilung der standardisierten Blocksumme.
Links: ziemlich symmetrisch bei $p = 0.4$. Rechts: rechtsschief (linkssteil) bei $p = 0.2$.

3.2 Beschreiben der Erzeugung von Textblöcken durch die Binomialverteilung

Bemerkung zur Simulation: Die Variation binärer Daten (0, 1) für eine Stichprobe vom Umfang 1.000 beträgt rund 0.03; d.h., eine Wahrscheinlichkeit kann mit dieser Präzision geschätzt werden, aber eben nicht genauer (jedenfalls, wenn wir 95-%-Konfidenzintervalle verwenden). Das bedeutet, unsere 1.000 Daten sollten nicht überinterpretiert werden, da es zusätzlich eben diese Quelle zufälliger Variation gibt. Abweichungen im Simulationsszenario können durch die geringe Präzision der Simulation oder durch die schlechte Anpassung an die Standardnormalverteilungskurve verursacht werden. Wir werden den Effekt der Simulation dadurch ausschalten, dass wir auf die Binomialverteilung zurückgreifen, welche die Texterzeugung modelliert. Diese neue Methode wird uns die Verbesserung der Anpassung in Abhängigkeit von der Länge des Textes ohne das „Rauschen“ der Simulation sehen lassen.

Wenn Text durch Zufallszahlen erzeugt wird, welche den Wert 1 mit Wahrscheinlichkeit p und 0 mit $1-p$ annehmen (und die Zufallszahlen verhalten sich so, als ob sie unabhängig wären), dann sind die einzelnen Zeichen im ersten Block der Länge 20 Zufallsvariablen $X_{1,1}, X_{1,1}, ..., X_{1,20}$ (der erste Index bezieht sich auf den Block, der zweite auf die Position des Zeichens innerhalb des Blocks) und die Blocksumme $B_1 = X_{1,1} + X_{1,1} + ... + X_{1,20}$ folgt einer Binomialverteilung mit $n = 20$ und p. Statt weiter Daten für die Blöcke zu simulieren, werden wir die potentiellen Ausgänge durch Wahrscheinlichkeiten aus dieser Binomialverteilung beschreiben. Diese Wahrscheinlichkeiten können als idealisierte relative Häufigkeiten aufgefasst werden. Wenn wir die Situation in Block i beschreiben, dann haben

wir eine analoge Situation: die Blocksumme $B_i = X_{i,1} + X_{i,1} + ... + X_{i,20}$ folgt derselben Binomialverteilung. Einerseits können Erwartungswert und Standardabweichung der Blocksumme aus allen Daten geschätzt werden, andererseits können diese Kennziffern für zukünftige Daten aus der Binomialverteilung vorhergesagt werden: $\mu = n \cdot p$ und $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.

Das Häufigkeitspolygon beschreibt die Verteilung der standardisierten Blocksummen; diese Daten werden durch die standardisierten Zufallsvariablen $\frac{B - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$ erzeugt (wir lassen den Index für die Blocknummer jetzt weg). Die gute Übereinstimmung mit der Standardnormalkurve lässt auch die Verteilung von B (eine Binomialverteilung) durch die Normalverteilung approximieren (die Rücktransformation ändert die Normalverteilung nicht): $B(n, p) \approx N(\mu = n \cdot p, \sigma = \sqrt{n \cdot p \cdot (1-p)})$. Wir haben auch herausgefunden, dass die Anpassung für p = 0.4 besser ist als für 0.2. Wir werden nun verschiedene Binomialverteilungen systematisch mit der entsprechenden Normalverteilung vergleichen. Jetzt werden wir nicht mehr standardisieren, sondern auf der Originalskala der Blocksummen bleiben.

3.3 Verschiedene Diagramme zum Darstellen einer diskreten Verteilung

Es gibt verschiedene Graphen, um diskrete Verteilungen darzustellen. Alle haben ihre Vorzüge und Nachteile. Wir werden „Schattengraphen“ verwenden, weil sie eine Flächeninterpretation stützen, was dann besonders wichtig wird, wenn wir eine diskrete mit einer stetigen Verteilung vergleichen.

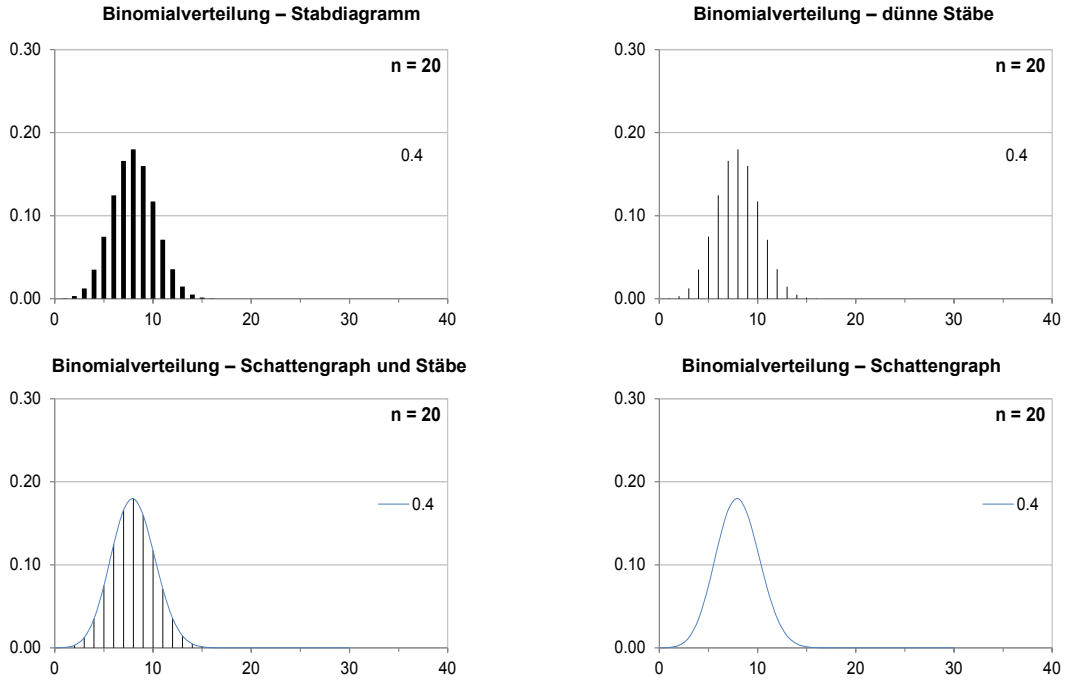

Abb 10: Stäbe, dünne Stäbe und Schattengraphen zur Darstellung der Binomialverteilung.

Für gewöhnlich wird die Binomialverteilung durch ein dickes Stabdiagramm dargestellt (Abb. 10); ein dünnes Stabdiagramm verdeutlicht, dass nur die einzelnen Punkte 0, 1, …, n von null verschiedene Wahrscheinlichkeiten haben; die Breite der Stäbe hat keinerlei Funktion. Für den Vergleich einer diskreten mit einer stetigen Verteilung wird die Fläche zum Schlüssel der visuellen Darstellung. Daher ersetzen wir die dünnen Stäbe durch eine Schattenlinie, welche die Höhen der Stäbe verbindet, und entfernen die Stäbe. Wir vergleichen diese Schattengraphen mit der Dichte der Normalverteilungen (nicht der Standardnormalkurve sondern in der Originalskala der Summen).

3.4 Analyse von künstlichem Text durch Inspizieren von Binomialverteilungen

Im Folgenden werden wir – statt Texte zu erzeugen – unser Wissen über die Binomialverteilung ausnützen. Diese beschreibt ja die theoretische Verteilung der Blocksummen. Die Analyse wird jetzt nicht auf standardisierte Blocksummen sondern auf Originalwerte erstreckt. Die Verteilungen werden durch „Schattengraphen“ dargestellt; wir variieren die Wahrscheinlichkeit p für das Zeichen 1 systematisch. Setzen wir für p kleine Werte ein (also etwa 0.1, 0.2 und 0.3), so sehen wir eine deutliche Schiefe und eine schlechte Anpassung der Normalverteilung (die Schiefe wird durch die Anpassung noch deutlicher sichtbar (Abb. 11), jedenfalls für n = 20. Ganz gleich sieht die Situation für große Werte von p aus. Um einiges besser passt die Normalverteilung für mittlere Werte (p = 0.4, 0.5 und 0.6), die Verteilungen sind ziemlich symmetrisch. Dazu zeigen wir kein Diagramm. Wir könnten den Vergleich mit n = 40 wiederholen und würden sehen, dass die Schiefe herausgenommen wird. Wir zeigen gleich die Diagramme für n = 100, welche eine frappante Übereinstimmung der Schatten der Binomialverteilung mit der entsprechenden Normalverteilung zeigen (Abb. 12; das illustriert auch die Brauchbarkeit der üblichen Faustregel $n \cdot p \cdot (1-p) > 9$ für die Zulässigkeit der Approximation).

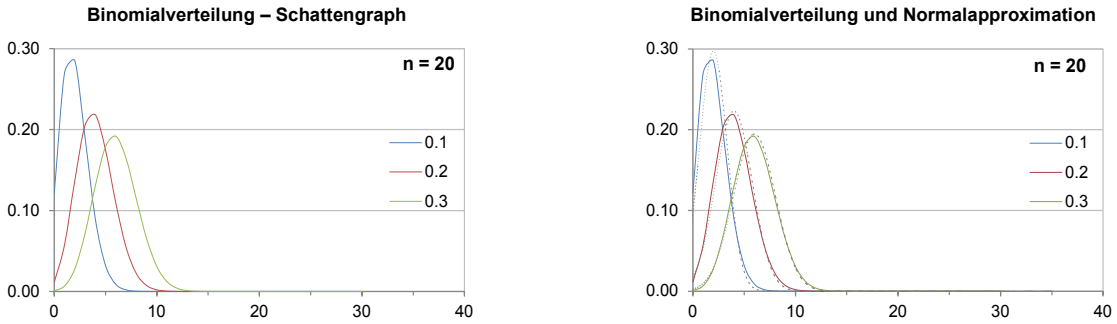

Abb. 11: Links: Beurteilung der Gestalt der Binomialverteilungen mit n = 20.
Rechts: Vergleich mit der Normalverteilung (strichlierte Kurven).

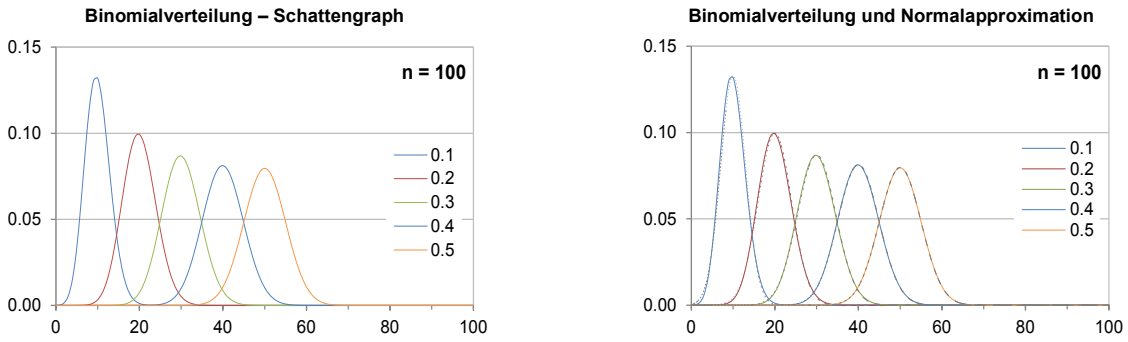

Abb. 12: Links: Beurteilung der Gestalt der Verteilung für Blocklänge 100 für p von 0.1 bis 0.5. Rechts: Die Normalverteilung (strichlierte Kurven) passt für p = 0.3 bis 0.5 so gut, dass sie fast mit dem Schatten der Binomialverteilung zusammenfällt.

Die Verbesserung der Anpassung durch eine Normalverteilung mit der größeren Länge der Blöcke (von 20 über 40 auf nun 100) gibt einen qualitativen Eindruck, dass die Anpassung sich im Sinne eines mathematischen Grenzwerttheorems immer verbessern sollte (der ZGS). Wenn die Blocklänge n über alle Maßen ansteigt („nach unendlich strebt“), dann sollte die Normalverteilungskurve als Grenzfunktion erscheinen. Es scheint klar, dass man sich aus mathematischer Sicht auf die standardisierten Blocksummen anstelle der Blocksummen beziehen muss. Für die Erzeugung von binärem Text haben die Blocksummen einen Erwartungswert von $n \cdot p$ und eine Standardabweichung von $\sqrt{n \cdot p \cdot (1-p)}$; beide Werte wachsen unbeschränkt, sodass im Grenzwert gar keine Wahrscheinlichkeitsverteilung existiert (das Grenzmaß ist das Nullmaß). Die Verteilung wird nach rechts verschoben und immer breiter bis sie ganz „einebnet“. Das ist auch der Grund für die anfänglich merkwürdige Standardisierung, welche in der Textanalyse verwendet wurde.

3.5 Kopf minus Zahl – Analyse eines Spiels anstelle von Texten

Wir spielen nun Münzwerfen und untersuchen den Saldo der Anzahl von Kopf und Zahl. Wir erzeugen den „Text“ nun durch ein Experiment, das nur zwei Zeichen hat: 1 für Kopf und −1 für Zahl. Auch hier führen wir Blöcke ein, d.h., wir fassen 20 Zeichen zu einem Block zusammen, und berechnen die Blocksumme, welche den Saldo eines Spielers darstellt, der gegen das Casino auf Kopf setzt; jedenfalls, wenn das Spiel symmetrisch wäre und der Spieler jeweils 1 € gewinnt oder verliert, je nachdem, ob Kopf erscheint oder ausbleibt. Wir sind an der Verteilung des Saldos des Spielers nach Beendigung eines Blockes interessiert. Um diese Verteilung zu schätzen, könnten wir das Spiel simulieren; wir bestimmen sie aus der Binomialverteilung mit n als der Blocklänge und $p = 0.5$ für eine ideale Münze.

Die Bandbreite der Summe steigt von 41 (Summe variiert von −20 bis 20) auf 81 und schließlich auf 201 (für $n = 20, 40, 100$) an. Das Diagramm wird breiter: sichtbare Stäbe (Wahrscheinlichkeit größer als ca. 0.05) von ±12 auf ±18 und schließlich auf ±24. Das Ergebnis eines Experiments ist die *Summe* von n Würfen, sodass wir erwarten, dass die standardisierte Summe von Kopf minus Zahl ungefähr einer Normalverteilung folgt. Der systematische Fehler durch Ersetzen von Stäben durch die Fläche unter einer stetigen Funktion wird kleiner, da durch die Standardisierung die Lücken kleiner werden (diesen Sachverhalt könnte man bei der Einführung der Integration ausnützen, siehe Engert, 2015).

Der Wertebereich der Zufallsvariablen Kopf minus Zahl wird auf ca. ±5 *reskaliert*. Wenn n wächst, besagt der ZGS, dass schlussendlich (ein Gedankenexperiment, welches in der Realität nie erreicht werden kann, da wir ja nur endlich lange spielen) die Verteilung der standardisierten Zufallsvariablen die Standardnormalkurve erreicht. Diese Grenzaussage stützt, dass man die Verteilung von Kopf minus Zahl (auf der ursprünglichen Skala) auch durch eine Normalverteilung approximieren kann. Man erhält die Originalskala aus den standardisierten Werten durch eine lineare Transformation (Skalierung und Verschiebung) zurück, was die Gestalt einer Normalverteilung invariant belässt (nur Erwartungswert und Standardabweichung werden entsprechend der Transformation angepasst).

In Abb. 13 sehen wir, dass die Verteilung von Kopf minus Zahl um die Null zentriert ist. Die Breite jedoch nimmt (ohne Schranke) zu. Es gibt keine Grenzverteilung für die Zufallsvariable Kopf minus Zahl. Eine Grenzverteilung taucht nur für die standardisierte Variable auf; d.h., wir haben zuerst den Erwartungswert (hier 0) abzuziehen und dann durch die Standardabweichung zu dividieren.

In dieser Weise verhilft uns der ZGS, auch *endliche Summen* der untersuchten Zufallsvariablen (Summe der einzelnen Würfe) zu approximieren. Da auch die Zufallsvariable *Durchschnitt* (Mittelwert) der Daten eine Reskalierung darstellt, bekommen wir eine Rechtfertigung, ihre Verteilung durch eine Normalverteilung zu approximieren. Aus praktischen Erwägungen sind wir weder an standardisierten noch an originalen Summen interessiert, wir zielen auf den Durchschnitt ab. Dieser hilft uns nämlich, das Mittel der Population zu schätzen, aus der die einzelnen Variablen ihre Werte zufällig entnehmen. Diese Population wird oft als endlich gedacht, aber in der Mathematik können wir sie auch als Prozess auffassen: ein Prozess des Münzwerfens etwa, welcher durch eine Zufallsvariable (welche die Werte 1 oder −1 annimmt, je nachdem, ob Kopf oder Zahl eintritt) modelliert wird.

Eine interessante Folgerung wird aus Varianten des Münzwerfens sichtbar. Mit einer verzerrten Münze ($p = 0.4$) mag der Spieler nach 20 oder sogar 40 Spielen einen positiven Saldo (Gewinn) haben, aber wir sehen, dass die Chancen dafür bei 100 Spielen drastisch abnehmen (Abb. 14). Das Risiko (die Wahrscheinlichkeit) für hohe Verluste dagegen ist erheblich angestiegen. Diese Eigenheiten werden noch markanter, wenn man die Zahl der Spiele erhöht. Casinos bieten im Normalfall weniger verzerrte Spiele an, in denen der Spieler längere Zeit durchaus gewinnen kann; aber die Struktur ist immer dieselbe: auf lange Sicht wird der Spieler alles verlieren. Die Chancen für eine einfache Wette im Roulette sind 0.4865 (18/37). Das hält die Spieler am Spielen, weil sie glauben, dass sie ihr eigenes Gewinnsystem gefunden haben (Abb. 15; die Gewinnzone hat hier kaum weniger Wahrscheinlichkeit als die Verlustzone, was beim ersten Hinsehen echt überrascht).

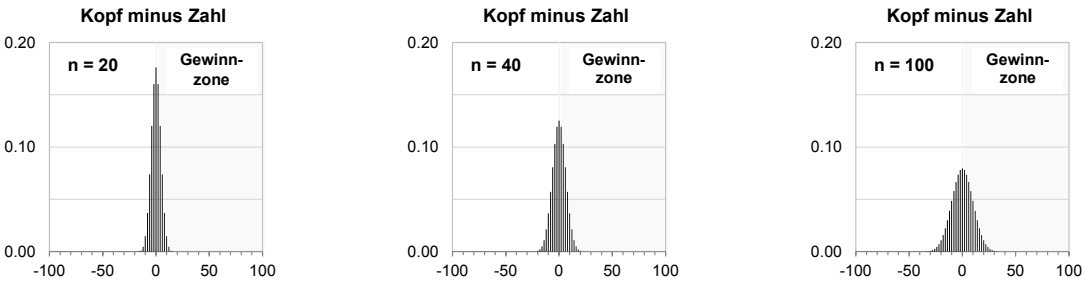

Abb. 13: Saldo der Anzahl der Köpfe und Zahlen für eine faire Münze nach *n* Versuchen.

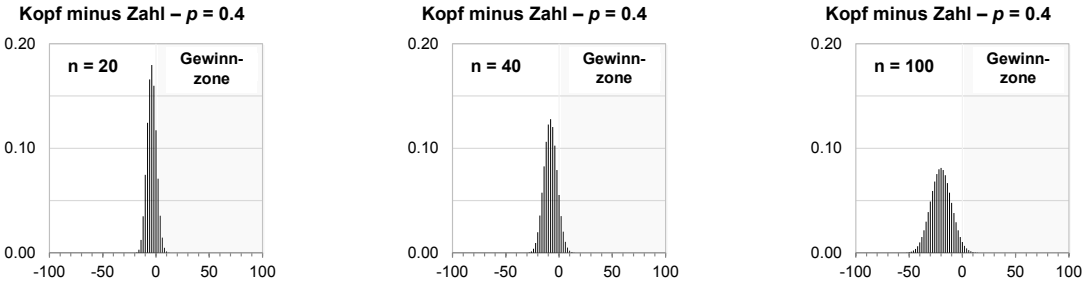

Abb. 14: Saldo der Zahlungen im Fall einer verzerrten Münze (p = 0.4).

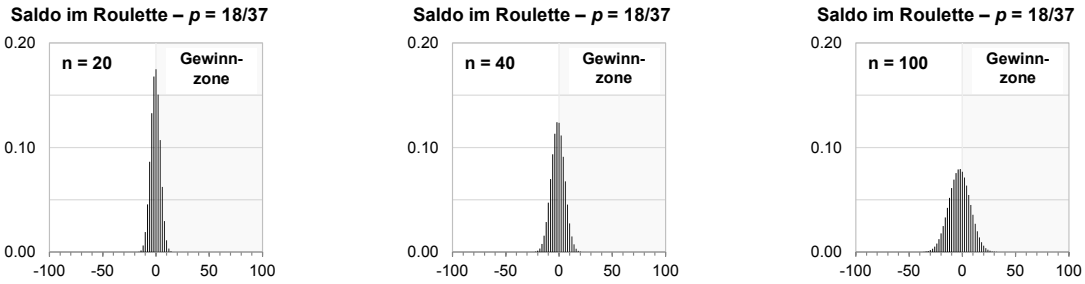

Abb. 15: Saldo der Zahlungen im Roulette beim Setzen auf Pair / Impair oder Rouge / Noir.

Wir haben das Spiel Kopf minus Zahl nicht nur deswegen eingeführt, um die schlechten Aussichten für notorische Spieler zu demonstrieren, sondern weil der ZGS in diesem Spezialfall mit relativ einfachen Hilfsmitteln beweisbar wird, jedenfalls für wirklich an Mathematik interessierte Schüler. Simulationsstudien liefern nur eingeschränkte empirische Evidenz, dass und unter welchen Umständen ein solches Theorem gelten sollte. Deshalb mag es gut sein, auch ein mathematisches Argument parat zu haben. Simulation kann ja ein mathematisches Argument nicht ersetzen; und es kann auch zu Verwirrung führen, weil wir ja kein Experiment unendlich oft durchführen können. Der Beweis (siehe Kusolitsch, 2001) folgt ziemlich genau dem Weg von de Moivre, der als Erster einen Term für die Normalverteilung eingeführt hat, als er Binomialwahrscheinlichkeiten im Spiel Kopf minus Zahl approximiert hat (LeCam, 1986). Moivre untersuchte die absoluten Werte dieser Zufallsvariablen und wendete die Stirling-Formel für n-Faktorielle an, um die harmonische Reihe in diesem Beweis zu approximieren.

4. Eine formalere Beschreibung der ursprünglichen Aufgabe

Für die ursprüngliche Aufgabe mit dem Text mit allen Zeichen und den Blocksummen können wir die Situation analog zu den Überlegungen mit binärem Text und der Binomialverteilung umformulieren.

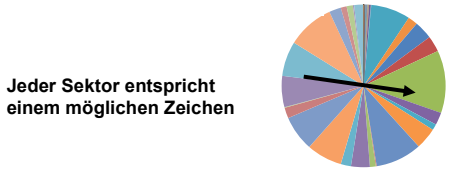

Abb. 16: Drehen des Glücksrads zwanzig Mal erzeugt einen Block (eine Stichprobe) der Länge 20.

In jedem Block wird die Summe durch ein allgemeineres Glücksrad „erzeugt“. Die Sektoren des Glücksrads entsprechen den im Text verwendeten Zeichen, die Fläche der Sektoren entspricht der Häufigkeit des Zeichens im Text (Abb. 18).

4.1 Blocksummen als Zufallsvariable und ihre Verteilung

Man erhält die Blocksumme durch Addieren der Ergebnisse von 20 hintereinander ausgeführten Drehungen eines allgemeinen Glücksrads (wie in Abb. 16), was zur Zufallsvariablen $B_{20} = X_1 + X_2 + ... + X_{20}$ führt. Wiederum wird die Approximation der standardisierten Blocksummen durch eine Standardnormalkurve durch den Mittelwert und die Standardabweichung des Glücksrads ausgedrückt; man beachte, dass wir den Zeichen numerische Codes zugeordnet haben. Die Verteilung für eine beliebige standardisierte Blocksumme (wir haben dies am Häufigkeitspolygon gesehen) ist näherungsweise die Standardnormalkurve:

$$\frac{B_{20} - \mu_{20}}{\sigma_{20}} \approx N(0,1).$$

Wir könnten ebenso feststellen, dass $B_{20} \approx N(\mu_{20}, \sigma_{20})$, indem wir unsere standardisierten Daten in die Originalskala rücktransformieren. Es bleibt noch zu untersuchen, wie Mittelwert und Standardabweichung für Blöcke der Länge 20 mit Mittelwert und Standardabweichung des Glücksrads (das beschreibt, wie die einzelnen Zeichen erzeugt werden) zusammenhängen: $\mu_{20} = 20 \cdot \mu$ und $\sigma_{20} = \sqrt{20} \cdot \sigma$ (dazu gibt es theoretische Argumente; oder wir prüfen an Daten, ob das gelten kann).

4.2 Der zentrale Grenzwertsatz (ZGS)

Der ZGS kann nun so formuliert werden: Gegeben seien *n* unabhängige Zufallsvariablen $X_1, X_2, ..., X_n$, welche alle dieselbe Verteilung haben wie *X*; *X* hat dabei Erwartungswert μ und eine endliche Standardabweichung σ. Für die Summe B_n definieren wir die standardisierte Zufallsvariable $\widetilde{B}_n = \frac{B_n - \mu_n}{\sigma_n}$. Die Verteilungsfunktion von $\widetilde{B}_n$ ist gegeben durch $F_n(z) = P(\widetilde{B}_n \le z)$. Die Verteilungsfunktion der Standardnormalverteilung (mit Erwartungswert 0 und Standardabweichung 1) sei wie üblich mit Φ notiert. Unter diesen Voraussetzungen gilt folgende Grenzwertaussage:

$$\lim_{n \to \infty} F_n(z) = \Phi(z) \text{ für alle reellen Zahlen } z.$$

Wir übersetzen das in den Kontext der Textanalyse: *X* ist der generische Term für die Erzeugung eines beliebigen Zeichens im Text und kann als „Glücksrad“ gedacht werden. X_2 ist dann die zweite Drehung und beschreibt, wie das zweite Zeichen erzeugt und dann ein numerischer Code (etwa ASCII) zugeordnet wird. Für einen Block der Länge *n* erzeugen wir *n* Zeichen. Die verschiedenen Drehungen werden intuitiv als unabhängige Versuche aufgefasst, was mathematisch der Unabhängigkeit der Zufallsvariablen $X_1, X_2, ..., X_n$ entspricht. Die Zufallsvariable B_n beschreibt, wie die Blocksumme aus den Werten der einzelnen Zeichen entsteht. Aus den Daten b_n von vielen (1.000) Textblöcken haben wir Mittelwert und Standardabweichung der Blocksumme geschätzt: $\mu_n \approx \bar{x}_{b_n}$ und $\sigma_n \approx s_{b_n}$. Daraufhin haben wir die standardisierten Blocksummen $\widetilde{b}_n = \frac{b_n - \bar{x}_{b_n}}{s_{b_n}}$ gebildet, welche Daten für die standardisierte Zufallsvariable $\widetilde{B}_n$ darstellen. Wir haben die Verteilung der standardisierten Blocksummen untersucht und dabei eine gute Anpassung an die Standardnormalkurve gefunden.

In natürlichen Texten ist die Unabhängigkeit bei aufeinanderfolgenden Zeichen eigentlich verletzt; wir haben daher versucht, durch zufälliges Umordnen der Zeichen die Unabhängigkeit hineinzubringen. Tatsächlich passt die Standardnormalverteilung dadurch um einiges besser. Wenn die Blocklänge n über alle Maßen anwächst (ein Gedankenexperiment, angelehnt an den mathematischen Sachverhalt einer Grenzwertaussage; empirisch können wir das kaum antizipieren), dann nähert sich die untersuchte Verteilung der Standardnormalkurve. Dies entspricht der Aussage des ZGS in seiner einfachsten Version (LeCam, 1986, beschreibt die spannende Geschichte dieses Theorems und seiner Verallgemeinerungen, welche Meilensteine auf dem Weg zur Axiomatisierung von Wahrscheinlichkeit darstellen; etwa müssen die Summanden gar nicht mehr dieselbe Verteilung haben).

Wir haben Häufigkeitspolygone für $n = 20$, 40 und 100 (letztere nur bei artifiziellem Text) untersucht und gesehen, dass sie der Standardnormalkurve sehr nahe kommen (wir könnten stattdessen auch Histogramme untersuchen). Unsere Ergebnisse liefern empirische Evidenz für den ZGS. Aus dem ZGS erhalten wir eine Rechtfertigung, bei endlicher Blocklänge die Verteilung der standardisierten Blocksumme zu approximieren. Wenn die Verteilungen konvergieren, so muss man irgendwann nahe genug beim Grenzwert sein. Es bleibt die Frage, wie groß n sein muss, damit die Approximation ausreicht. Wir haben gute Übereinstimmung schon bei $n = 20$ gefunden. Die Diskussion um die Mindestgröße von n hängt eng mit der Erzeugung der einzelnen Zeichen zusammen (mit der Verteilung der Werte, welche diesen Zeichen zugeordnet werden). Blocklängen von $n = 100$ reichten für $p = 0.1$ nicht ganz aus in der artifiziellen binären Texterzeugung (obwohl die Verteilung nur mehr leicht schief war). Für kleinere Werte von p braucht man noch längere Blöcke, damit die Approximation gut wird.

4.3 Folgerungen aus dem ZGS – Normalapproximation für Summen und Mittelwerte

Wenn wir eine Rechtfertigung für die Approximation der Verteilung von standardisierten Blocksummen durch eine Standardnormalkurve haben, können wir sie auch nutzen, um die Blocksummen in der Originalskala durch eine Normalverteilung zu approximieren. Wir müssen nur die Parameter von 0 und 1 der Verschiebung und dem Skalierungsfaktor anpassen, mit denen wir die Blocksummen standardisiert haben; d.h., die Parameter der approximierenden Normalverteilung sind $\mu_n = n \cdot \mu$ und $\sigma_n = \sqrt{n} \cdot \sigma$. Analog stellt das Blockmittel $M_n = B_n / n$ eine Reskalierung dar, welche auch mit einer Normalverteilung, jetzt mit den Parametern μ und $\sigma / \sqrt{n}$, approximiert werden kann. Wir werden diese Zusammenhänge zwischen den Parametern für die statistische Inferenz noch benötigen. Diese Beziehungen zwischen Mittelwert und Standardabweichung für die verschiedenen Statistiken können aus Daten geschätzt werden. Bei der künstlichen Erzeugung von binärem Text können wir auch Wissen über die Binomialverteilung einbringen; ein Wissen, das mit den obigen Gleichungen vereinbar ist. Wir könnten auch intuitive Argumente anführen, welche diese Beziehungen stützen, jedenfalls in einfachen Spezialfällen (siehe Borovcnik, 2001 und 2011). Ein allgemeiner Beweis erfordert mehr Mathematik, weshalb man vielleicht geneigt sein kann, die Beziehungen durch Analyse von Daten aus Computersimulationen zu stützen.

Wie man in Abb. 17 sehen kann, gibt es keinen ZGS für Summen oder für Mittelwerte von Blöcken. Während die Verteilung der Summen dazu tendiert, immer größer und flacher zu werden, bis keine Wahrscheinlichkeitsverteilung mehr übrig bleibt (das Nullmaß), zieht sich die Verteilung von Mittelwerten (nach dem GGZ) zusammen auf eine Achse, die gerade dem Erwartungswert des Glücksrads entspricht, das die einzelnen Daten erzeugt (das ist eine degenerierte Ein-Punkt-Verteilung). Da wir aber eine Rechtfertigung für die Normalapproximation von standardisierten Blocksummen haben, können wir dasselbe Argument auch für die Blocksummen und Blockmittelwerte verwenden, weil diese aus den standardisierten Summen durch eine lineare Transformation hervorgehen. Diese Reskalierung verändert die Normalverteilung selbst nicht, nur ihre Parameter müssen angepasst werden.

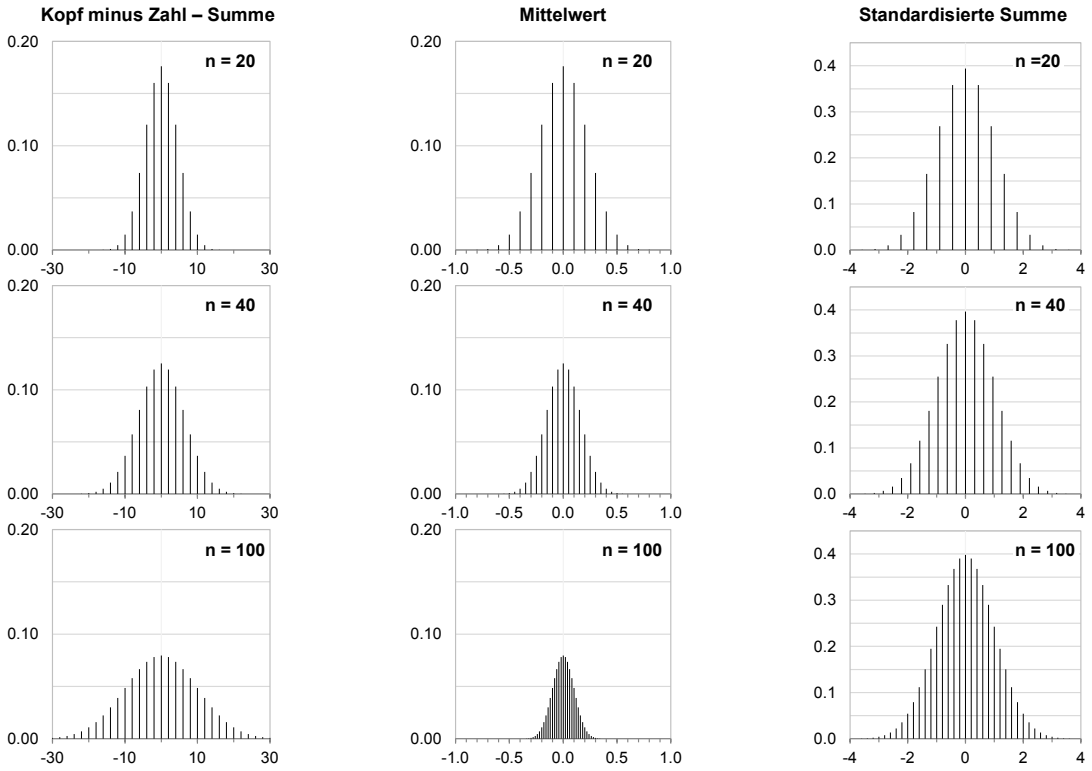

Abb. 17: Die Summe divergiert (linke Spalte) – der Durchschnitt konvergiert zu einem einzelnen Punkt (mittlere Spalte) – die standardisierte Summe konvergiert in Verteilung zur Standardnormal-Dichte (rechte Spalte).

5. Stichproben und Populationen – statistische Inferenz

Wir haben bislang faktischen Text oder eine Methode, Text zu erzeugen, untersucht. Dazu haben wir den Text in Blöcke bestimmter Länge unterteilt und die Verteilung der Blocksumme untersucht. Wir sehen uns nun diese Analyse mit statistischen Augen an und betrachten dazu den Text als Population und die Textblöcke als Stichproben, aus denen Information über den Text extrahiert werden soll.

5.1 Interpretation der Textanalyse in Form von Stichproben und Population

Wir interpretieren den Text als Population um, welche untersucht werden soll. Ist der Text endlich, sprechen wir von endlichen Populationen; wird ein Prozess zur Erzeugung von Text untersucht, so sprechen wir von unendlichen Populationen. Die Verteilung der ASCII-Codes im ganzen Text wird damit zur Population, aus welcher wir unsere Stichproben ziehen. Analog werden wir die Methode zur Erzeugung von Text (ein Glücksrad) als Population ansprechen. Wir können neben binären auch kompliziertere Glücksräder verwenden, um Text zu erzeugen.

Die Textblöcke werden in dieser Sicht zu Stichproben. Die zufällige Umordnung des Textes hat die Anpassung der Standardnormalkurve verbessert. Zufällige Blöcke oder Stichproben garantieren, dass einzelne Zeichen ausgetauscht werden können, ohne dass die grundsätzlichen Eigenschaften des Textes hinsichtlich seiner Zusammensetzung (nicht in der Abfolge!) verändert werden. In der beurteilenden Statistik führen wir Methoden zur Verallgemeinerung von Information aus den Daten einer Stichprobe auf die Population ein. In der Textanalyse könnten wir am Mittelwert der Population (alle ASCII-Codes mit ihrer Häufigkeit, oder $p = 0.4$ in der artifiziellen Generierung von Text) interessiert sein. Basisinformation ist ein Textblock. In Abb. 18 liefert die Verteilung der ASCII-Codes im Text (links) oder das Stabdiagramm (rechts) eine statische Sicht auf den Text: wie sind die Werte verteilt, die den Zeichen des Textes zugeordnet werden. Das Glücksrad repräsentiert dagegen eine dynamische Sicht auf dieselbe Population; durch Drehen können Textblöcke (Stichproben) erzeugt werden, welche

in sich Information über die Population tragen. Der Prozess der Erzeugung von Text wird zur Population, während die erzeugten Textblöcke nun zu Stichproben werden.

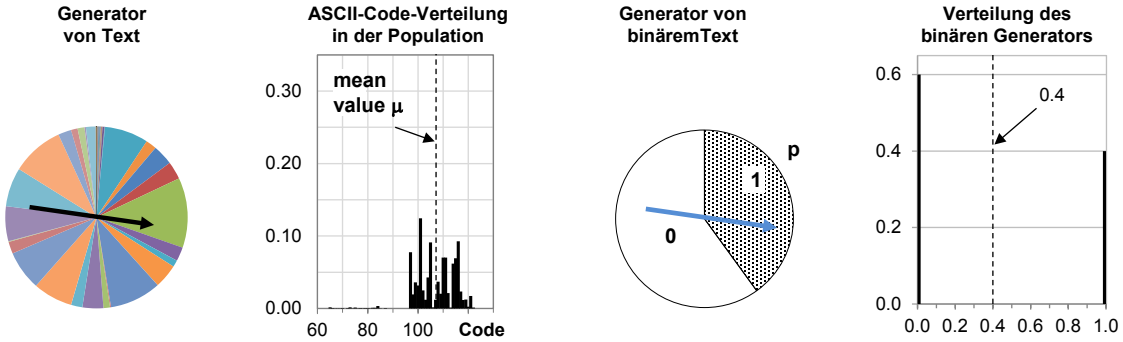

Abb. 18: Darstellung der Population – Erzeugendes Glücksrad und Stabdiagramm.

Im Allgemeinen haben wir nur *einen* Textblock, d.h., eine Stichprobe vom Umfang *n*. Um die Beziehungen zwischen dem Blockmittel (Mittelwert der Zeichen in der Stichprobe) und dem „Mittelwert" in der Population zu untersuchen, benutzen wir die *Verteilung* der untersuchten Statistik. Wie variiert die Blocksumme (der Mittelwert), wenn eine andere Stichprobe (ein anderer Block) untersucht wird? Wir haben die Verteilung der Blocksumme (der Summe aller Werte einer Stichprobe) untersucht. Der ZGS besagt, dass die standardisierte Blocksumme durch die Standardnormalkurve angepasst werden kann. Das rechtfertigt, sowohl die Verteilung der Blocksumme als auch die Block- oder Stichprobenmittelwerte durch eine reskalierte Normalverteilung zu approximieren.

5.2 Schätzen des Mittels einer Population aus Blöcken oder Stichproben

Statistische Inferenz hat mit dem Studium der Zusammenhänge zwischen dem erzeugenden Glücksrad und dem generierten Text zu tun. Wir zeigen, wie diese neue Sicht auf Population und Stichproben Schlüsse von einem einzelnen Wert (einem Mittelwert der Daten) in einer Stichprobe auf den Mittelwert der Population rechtfertigt. Das Zusammenziehen der Verteilung um den Mittelwert der Population entspricht dem GGZ (für Mittelwerte). Der ZGS stellt numerische Wahrscheinlichkeiten parat, mit denen bestimmte Schranken der Abweichung von Stichproben- und Populationsmittel überschritten werden, wenn man eine Stichprobe des Umfangs *n* hat. Wir zeigen zwei Folgen von Diagrammen (Abb. 19 und 20), eine für allgemeine Glücksräder (allgemeiner „Text") und eine für unsere Methode, binären Text zu generieren. Die Methode, Text zu erzeugen, wird zur Methode, Stichproben zu ziehen.

Abb. 19 zeigt, dass sich der Mittelwert der Population (des Glücksrads, welches den Text erzeugt) in den Mittelwerten der Textblöcke spiegelt: Mittelwerte einzelner Textblöcke streuen um eine Achse, die durch den Mittelwert der Population bezeichnet wird. Die Mittelwerte der Textblöcke liegen näher zu dieser Achse, wenn die Blocklänge zunimmt. Als ein Gedankenexperiment – diese Mittelwerte werden sich ganz auf diese Achse zusammenziehen, wenn die Länge ohne Schranke vergrößert wird.

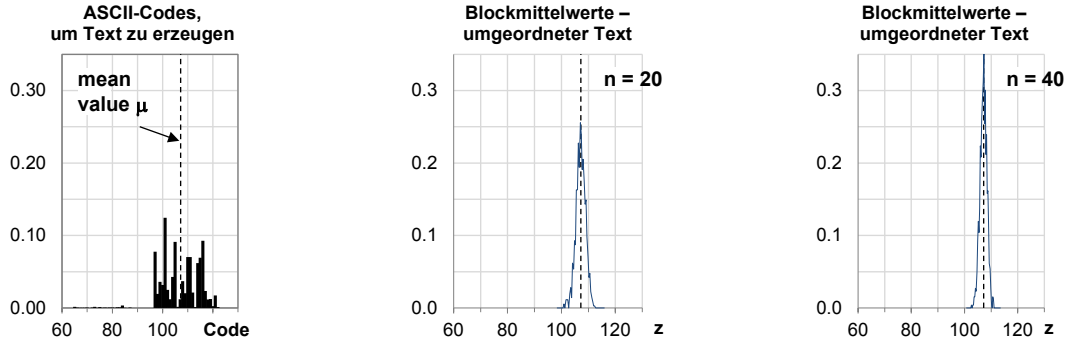

Abb. 19: Verteilung der ASCII-Codes in der Population, welche den Text erzeugt, und Verteilung der Durchschnitte in Böcken.

5.3 Schätzen eines Anteils aus dem „Mittel“ von Blöcken oder Stichproben

Für binären Text sind die Beziehungen zwischen erzeugendem Glücksrad und die dadurch erzeugten Textblöcke analog zum allgemeinen Glücksrad. Aus Abb. 20 sehen wir, wie die Verteilung der Block- oder Stichprobenmittelwerte enger um den Mittelwert der Population (dem Glücksrad) wird. Wieder können wir uns – in einem Gedankenexperiment – vorstellen, wie sich die Verteilung auf diese Achse zusammenzieht, wenn wir immer längere Blöcke (Stichproben) untersuchen.

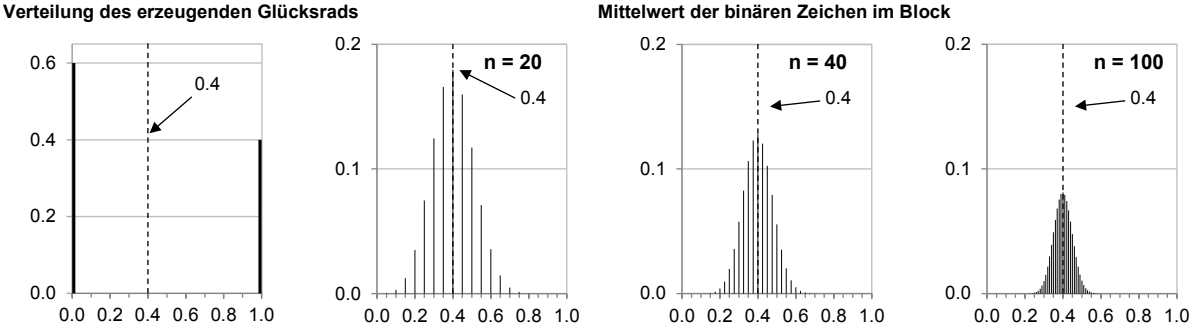

Abb. 20: Binärer Text – der Durchschnitt von Stichproben trägt Information über die Wahrscheinlichkeit der 1en.

5.4 Folgerungen aus dem ZGS und dem GGZ für Stichprobenziehungen

Der ZGS garantiert, dass Abweichungen zwischen Stichprobenmittel und Mittel der Population, welche eine gewisse Schranke überschreiten, mit der Normalverteilung berechnet werden können. In Verallgemeinerung ist jede Statistik annähernd normalverteilt, welche durch eine Summe von Werten einzelner Einheiten der Stichprobe berechnet wird (und das ist fast jede Statistik). Die Überlegungen in diesem Aufsatz erklären, warum die Normalverteilung für die statistische Inferenz so wichtig ist. Experimente zum GGZ findet man in Borovcnik (2001) oder in Borovcnik und Schenk (2012).

Der Autor dankt Herrn Reinhard Winkler für die kritische Durchsicht des Beitrags, welche die Lesbarkeit der Ideen erheblich verbessert hat.

Literatur

Borovcnik, M. (2001): Nützliche Gesetze über den Zufall – Experimente mit Excel. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft* (ÖMG), 33, 1-22.

Borovcnik, M. (2013): Bedingte Wahrscheinlichkeit – Ein Schlüssel zur Stochastik. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft* (ÖMG), 16, 1-18.

Borovcnik, M. (2015): Central theorems of probability theory and their impact on probabilistic intuitions. In: J. M. Contreras et al (Hsg.): *Didáctica de la Estadística, Probabilidad y Combinatoria 2*. Granada, 15-35.

Borovcnik, M. (2011): Key properties and central theorems in probability and statistics – corroborated by simulation and animation. *Selcuk Journal of Applied Mathematics, Special issue on Statistics*, 3-19.

Borovcnik, M.; & Schenk, M. (2012): Simulationen im Stochastik-Unterricht. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft* (ÖMG), 44, 1-16.

Engert, I. (2015): *Ein genetisch orientierter Lehrgang zur Wahrscheinlichkeitsrechnung*. Dissertation: Salzburg.

Kusolitsch, N. (2001). Bemerkungen zum zentralen Grenzverteilungssatz im Mittelschulunterricht. *Berichte des Instituts für Statistik*. Wien: TU Wien.

LeCam, L. (1986): The central limit theorem around 1935. *Statistical Science*, 1, 78-96.

Nemetz, T.; Simon, J.; Kusolitsch, N. (2002): Überzeugen statt Beweisen – der zentrale Grenzverteilungssatz im Gymnasialunterricht. *Stochastik in der Schule*, 22 (3), 4-7.

Verfasser

Manfred Borovcnik
Alpen-Adria-Universität Klagenfurt, Institut für Statistik
Universitätsstraße 65, 9020 Klagenfurt
manfred.borovcnik@uni-klu.ac.at