Central theorems of probability theory and their impact on probabilistic intuitions

Manfred Borovcnik

manfred.borovcnik@aau.at, Alpen-Adria University, Klagenfurt

Abstract

The Central Limit Theorem (CLT) substantiates the normal distribution, which becomes a key player in probability and statistics. A simplification of the mathematics is needed so that students can shape their intuitions on probability. The CLT justifies using the normal distribution as an approximation for random variables that are or that can be thought to be the sum of other random variables. Our key experiment has to do with text analysis from a statistical point of view. It is surprising and motivating for learners that we can predict the shape of the distribution, which is investigated. The considerations also motivate how the continuous standard normal distribution can be the limit of discrete distributions. Text analysis provides a natural context to discuss interrelations between samples and populations, which form the core of inferential statistics.

Keywords: Standardized sums; Normal approximation; Class experiment

1. Introduction

Probability is a difficult concept and there are many misleading intuitions. Unlike in geometry our perception has not been trained to improve our ideas as probability is not a physical property in the real world. Yet it is often equated to the relative frequencies of an event in a series of repeated experiments. In fact, there is a relation between the two concepts (if only such an experiment could be repeated under the same conditions) – though this relation is a bit more complicated. Some statisticians therefore prefer to speak of probability as a metaphor to communicate about a random situation, or they would state that probability is a virtual concept (like the Internet or computer games are virtual worlds).

Mathematically three groups of central theorems regulate what probability is and how we can interpret it. The one group of theorems is the laws of large numbers; the second is the group of Central Limit Theorems. The third is Bayes' theorem by which subjective probabilities converge to the relative frequencies. The first justifies that we interpret probability in terms of relative frequencies. The basic Law of Large Numbers is usually summarized as: the relative frequencies "converge" to the (possibly) unknown probability of the event under scrutiny. The second explains why we can describe the variation of a random variable by a normal distribution in quite a few cases (and becomes eminently important in statistical inference). The simplest case has become famous in the history of probability as the law of errors, which is a thought experiment: if a measurement error can be explained by a sum of elementary errors (each of them is not observable) then the resulting error (that can be observed) should follow a normal distribution. Bayes' theorem shows how we can improve qualitative knowledge by data.

The simplifying statement for the Law of Large Numbers is simply wrong and misleading but is has a true kernel. We could look more precisely at the mathematical theorem but this requires quite a lot of sophisticated arguments. The question is how to develop scenarios and formal signs (with accompanying pictures behind) that we can teach the topic and communicate

En J. M. Contreras, C. Batanero, J. D. Godino, G.R. Cañadas, P. Arteaga, E. Molina, M.M. Gea y M.M. López (Eds.), Didáctica de la Estadística, Probabilidad y Combinatoria, 2 (pp. 15-35). Granada, 2015.

its relevance, shape intuitions that comply with the mathematical background, and "revise" intuitions that are at least not helpful (if not wrong). How can we explain at an intuitive level, in which sense and under which conditions the relative frequencies do converge to the underlying probability? The simplifying statement for the Central Limit Theorem is simply wrong as the sum of the elementary errors cannot converge as it tends to get larger and larger if we add more elementary errors (even with an increasing variability). Again, the teaching challenge is to investigate various situations and observe a kind of divergence or convergence. A further challenge is to clarify the kind of convergence to the normal distribution and design situations where such knowledge would be helpful.

We will use simulations of random experiments and didactical animations of binomial distributions and investigate the "data" from various perspectives to support feasible ideas about the Central Limit Theorem, which will help to understand how the concept of probability may be used to extract information from data.

2. Analysis of a natural plain text

Text analysis and interpretation is a sophisticated discipline of linguistics. We will perceive text analysis in a "narrow" way. We attribute numerical codes to the signs of the text and analyse, amongst others, the frequency of the codes, or the distribution of the code sum of smaller cuts of the whole text.

The reader might remember times as child when someone tried a magic trick upon them starting with "think of two numbers between 1 and 10". Then the steps were to add the two numbers; subsequently to take the square of the sum; then to multiply the result by 9, take the square root of the intermediary number, subtract three times the second number, and divide the result by the first number. "You must have got a 3!" the person told us. We were amazed. How could this person know that number?

We will discuss an analogue experiment based on the "analysis" of texts. Instead of think of two numbers, we ask the test person to deliberately choose a text of a certain length. Instead of performing calculations with the chosen numbers, we ask to investigate the distribution of numbers that are attributed to blocks into which the text is partitioned. We cannot tell the exact distribution of the test person but predict that it looks similar to a standard normal curve.

2.1. The experiment

We follow a recommendation of Nemetz, Simon, and Kusolitsch (2002). Take a longer text, any text of your choice. Remove any blanks, special signs as periods, colons, semicolons, commas, numbers, and brackets from the text. Make sure you have exactly 20,000 signs left. Arrange the signs in one column of a spreadsheet (we will help you with that). Attribute a code number from 1 to 1000 to each of the possible signs (your choice). Separate all the signs into blocks of 20. Calculate the sum of the first 20-block, calculate the sum of numbers attributed to the signs of the second block, and repeat calculating the sum of all 1,000 blocks of 20 numbers (that are attributed to the signs of the block).

You end up with 1,000 block sums (1,000 data). Calculate the mean and the standard deviation of the block sums. After that obtain the standardized block sums, i.e., subtract from each block sum the mean and divide the result by the standard deviation. You have now 1,000 stand-

ardized block sums. I can tell that nearly all your standardized data are within the limits of −5 and 5 and if you find a histogram for your data, it will be close to the standard normal curve.

Tell two friends to join in the experiment. They should find their own attribution of numbers to the signs. Their final histogram will be quite similar to yours and to the standard normal curve. Repeat the experiment with 40,000 signs and build blocks of length 40. Your final histogram will even be closer to the standard normal curve as before. You can choose any other text you like. You can also perturb your signs in the text randomly (this is easily done by random numbers) and you will witness an even better fit of your histogram to the standard normal curve. How could we tell the result before? It is not a trick; as in the game of our childhood the result could be explained. However, the explanation goes beyond simple equations and has to do with the Central Limit Theorem. We will first show the progression of the game with a special text.

Table 1. Example of text coding

Signs	Codes	Nr.	Block nr.	Pos.in block	Signs	Codes	Nr.	Block nr.	Pos.in block
R	82	1	1	1	g	103	21	2	1
i	105	2	1	2	T	84	22	2	2
s	115	3	1	3	h	104	23	2	3
k	107	4	1	4	e	101	24	2	4
a	97	5	1	5	L	76	25	2	5
n	110	6	1	6	o	111	26	2	6
d	100	7	1	7	g	103	27	2	7
D	68	8	1	8	i	105	28	2	8
e	101	9	1	9	c	99	29	2	9
c	99	10	1	10	o	111	30	2	10
i	105	11	1	11	f	102	31	2	11
s	115	12	1	12	P	80	32	2	12
i	105	13	1	13	r	114	33	2	13
o	111	14	1	14	o	111	34	2	14
n	110	15	1	15	b	98	35	2	15
M	77	16	1	16	a	97	36	2	16
a	97	17	1	17	b	98	37	2	17
k	107	18	1	18	i	105	38	2	18
i	105	19	1	19	l	108	39	2	19
n	110	20	1	20	i	105	40	2	20

2.2. Specific steps of the analysis of the text

We used a recently published paper on risk and attributed the ASCII code to the signs. In Table 1 we show only the result of coding for the first two 20-blocks. We calculate the sum of these two blocks and get $b_1 = 2026$ and $b_2 = 2015$. In Table 2 (left) we show the block sums for the first twenty 20-blocks just to give a flavour of the variation. From all block sums we then calculate the mean and standard deviation and get (from our data, which are available from an Excel file) $\bar{b} = 2143.32$ and $s_b = 33.08$. The first standardized block sum now is obtained by $\frac{b_1 - \bar{b}}{s_b} = \frac{2026 - 2143.32}{33.08} = -3.5469$. We continue with the other blocks and obtain 1,000 standardized sums. The relative frequencies of the classes (−5, −4.8], (−4.8, −4.6], …, (4.8, 5] may be denoted by f_i; we calculate a data density (like a population density) by dividing by the width of

the classes (0.2) and draw a density polygon connecting the points (midpoint of class i, $f_i/0.2$) from our data on the standardized block sums (see Table 2, right side). We prefer a frequency polygon over a histogram as it gives a clearer interpretation of a *function* as we compare this density polygon to the standard normal curve. Again, we show only part of the frequency table.

Table 2. Computations in the analysis of the text

Block number	Block sum	Mean, sd	Standard-ized sum	Class $(e_{i-1}, e_i]$	Mid-point m_i	Frequency abs. n_i	Frequency rel. f_i	Density f_i /0.2	Density stdnormal
1	2026	2143.32	-3.5469	-5.0					
2	2015	33.08	-3.8794	-4.8	-4.9	0	0.000	0.000	0.000
3	2052		-2.7608	-4.6	-4.7	1	0.001	0.005	0.000
4	2077		-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

The frequency polygon in Figure 1 comes quite close to the standard normal curve. However, compared to our great "promise" the fit could be improved, especially in the centre, left and right of zero! This is caused by our text that has distinct sequences of signs.

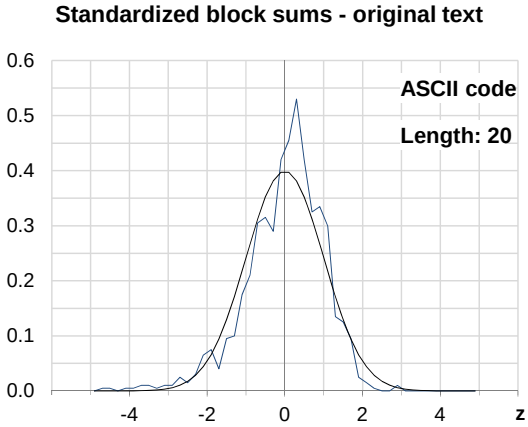

Figure 1. Frequency polygon (of data density) for the standardized block sums in the original text – based on the attribution of ASCII codes to single signs

2.3. Improve the fit of the standard normal curve by making the text more random

We will repeat the analysis with rearranging our original text by a random sequence. We can obtain this random reordering by perturbing the signs by random numbers (see below). We will show only the resulting frequency polygon; the random derangement in fact has increased the fit enormously (Figure 2).

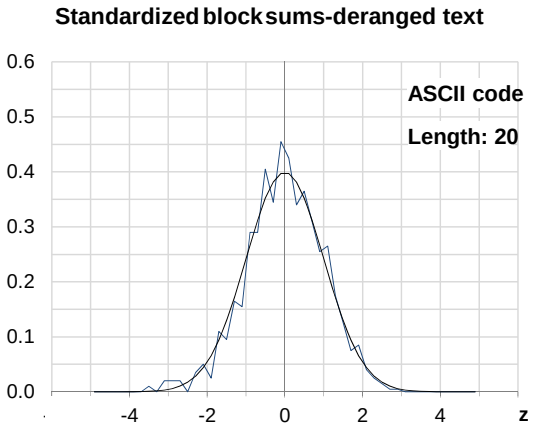

Figure 2. Frequency polygon (of data density) for the standardized block sums in the randomly re-arranged text – based on the attribution of ASCII codes to single signs

2.4. Inspection of the impact of the coding scheme

One might suspect that something has been done with the attribution of codes that "caused" the good fit to the standard normal curve. However, an inspection of the distribution of the assigned codes looks quite unusual and scattered (Figure 3). Nothing in it "resembles" a normal distribution. There are quite a few outliers in the range between 45 and 90, scattered unevenly over a long interval. It seems even more amazing that finally the normal curve fits so well.

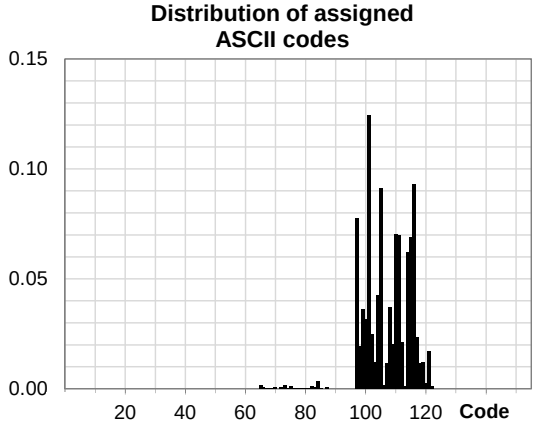

Figure 3. Distribution of the codes for the whole text of 20,000 signs for the ASCII code

We now have a look on the results of the other two persons who encoded the signs of the text differently. We suppose that one has ensuing numbers NR from 1 to 55, which is – compared to the ASCII code, a very compact encoding without gaps in between. For the other

person, we assume that 20,000 random numbers have been generated and ordered so that RD_i is the i-th smallest random number. If a sign has been encoded by NR = i then the random encoding would assign $i \cdot RD_i \cdot 10$ and take the integer part of this number. This method should ensure that the code is mainly established by randomness.

We investigate the frequencies of block sums (with block length 20) in the same way as earlier with the ASCII codes and finally get the following frequency polygons of the standardized block sums, which show roughly the same fit by a standard normal curve (Figure 4 left). For these two encoding systems we only show the polygon for the randomly rearranged text as we have noticed earlier that the single signs show a kind of slight dependence and the random order of the text fits much closer to the standard normal curve. In both cases there is a slight overrepresentation of the first interval left to 0 (see Figure 4; with the random code also the second interval is slightly overrepresented).

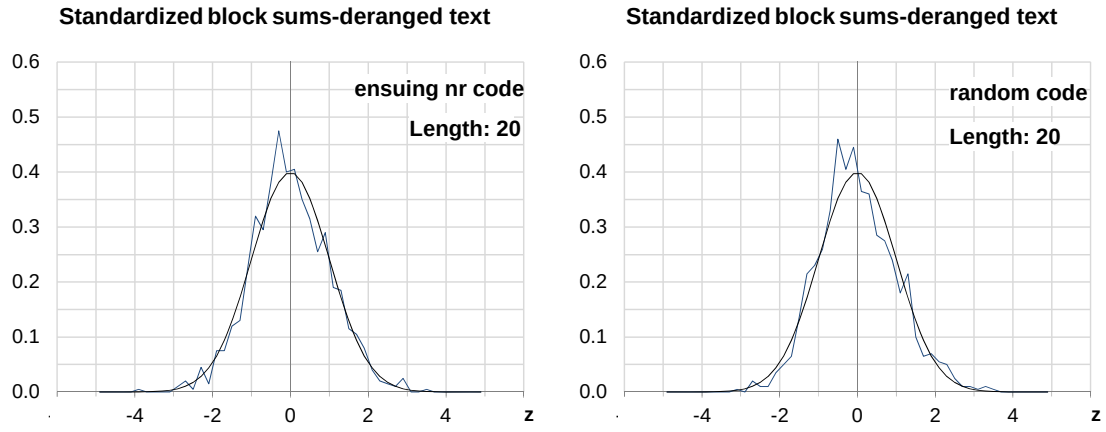

Figure 4. Frequency polygon (of data density) for standardized block sums in the randomly rearranged text. Left: based on ensuing numbers as codes; Right: based on random codes

The coding in Table 3 (only a part of it is shown) gives a flavour of the actual attribution of signs in the text to the codes. As with the ASCII code we might inspect the distribution of the single codes in the whole text. With the ensuing number code the distribution is compact but very uneven, the random assignment has a much greater variability (the first axis stretches from 0 to 500 as compared from 0 to 50 for the ensuing numbers, Figure 5).Yet the final result – the fit of the standard normal curve – is similar for both.

Table 3. Part of the coding table of the various systems used for our analysis

Sign	ASCII	Nr	Rand	Sign	ASCII	Nr	Rand
A	65	1	0	F	70	11	29
a	97	2	1	f	102	12	32
B	66	3	3	G	71	13	34
b	98	4	4	g	103	14	38
C	67	5	10	H	72	15	42
c	99	6	12	h	104	16	46
D	68	7	14	I	73	17	50
d	100	8	17	i	105	18	55
E	69	9	22	J	74	19	63
e	101	10	25	j	106	20	67

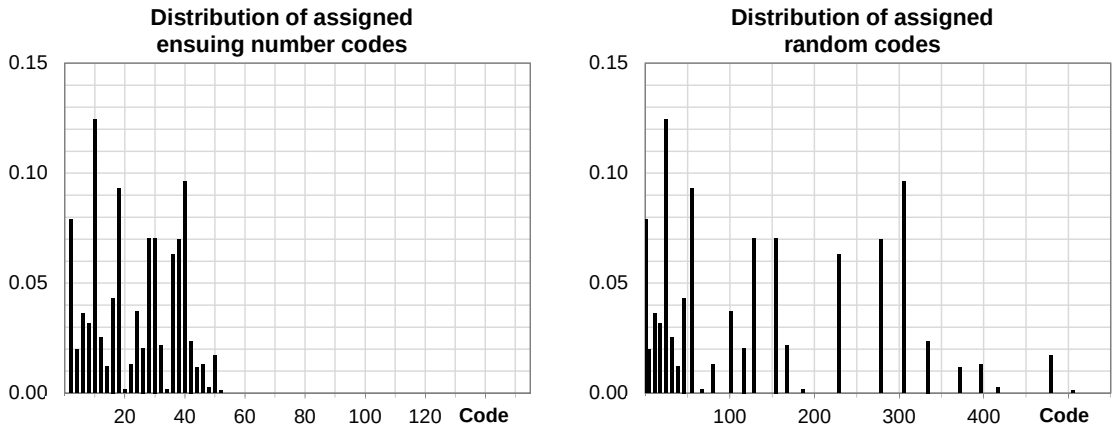

Figure 5. Distribution of the codes for the whole text with ensuing number and random code

2.5. Impact of text length

We still have to investigate the effect text length on the shape of the frequency distribution of block sums. While there is some improvement (Figure 6), the improvement expected from theory has not been totally fulfilled. That is due to specifities of text that do not only cause dependencies between ensuing signs (which should be removed by the rearrangement) but also restricts the letters in several of the longer blocks. If the text is on risk, risk, e.g., will be referred to quite often, etc. We will see that if we artificially generate text, the fit of the standard normal curve will considerably be increased by doubling the text length.

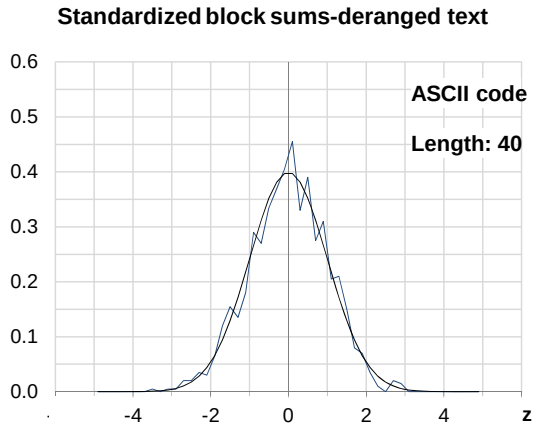

Figure 6. Frequency polygon (of data density) for the standardized block sums in the randomly re-arranged text – based on the attribution of ASCII codes to single signs – block length 40

3. Generating artificial texts with only two signs

Instead of using available texts, we will now generate our own text so that it fits more closely to the probabilistic assumptions. We will use only two signs and encode them by 0 and 1. The signs will be produced independently, which may be interpreted as if a wheel of chance with two sectors is spun several times (Figure 7). By twenty spins we generate one block of length

20. We then will repeat the procedure 1,000 times in order to imitate the text analysis from earlier sections.

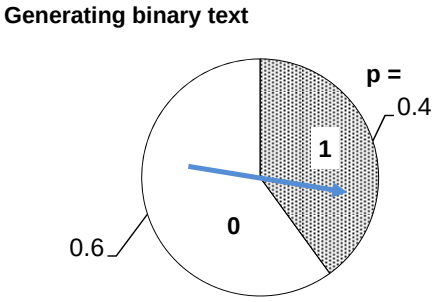

Figure 7. Generating a block of length 20 means spinning the wheel 20 times

3.1. Analysing artificial text

We generate a binary text randomly with 0's and 1's; first we will use $p = 0.4$ for sign 1. Then we proceed in our analysis as before. We calculate the block sum and standardize the values according to our 1,000 data that we generated all in all. The distribution of these standardized values is again displayed by a frequency polygon.

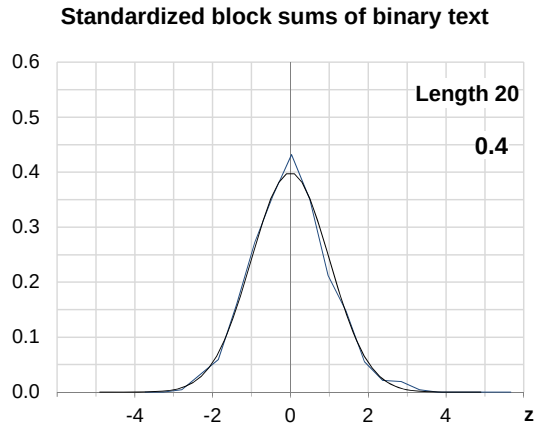

Figure 8. Binary text with 0 and 1 ($p = 0.4$ for sign 1) – standardized block sum – frequency polygon compared to standard normal curve

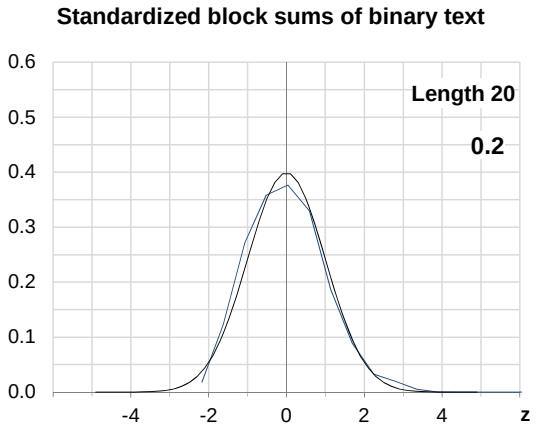

Figure 9. Binary text with 0 and 1 ($p = 0.2$ for sign 1) – standardized block sum – frequency polygon compared to standard normal curve

It is amazing how good the fit is (Figure 8). If we generate a text with a lower value of p (0.2) then the fit is not so well (Figure 9) but would increase again if the number of signs in the single blocks is increased. The polygon shows a systematic shift to the left as compared to the standard normal curve. We replicate the generation of text by simulation and we split the text of 40,000 signs now into blocks of length 40. To show also the "noise" of simulation, we display two frequency polygons with $p = 0.4$ and with $p = 0.2$. (Figure 10).

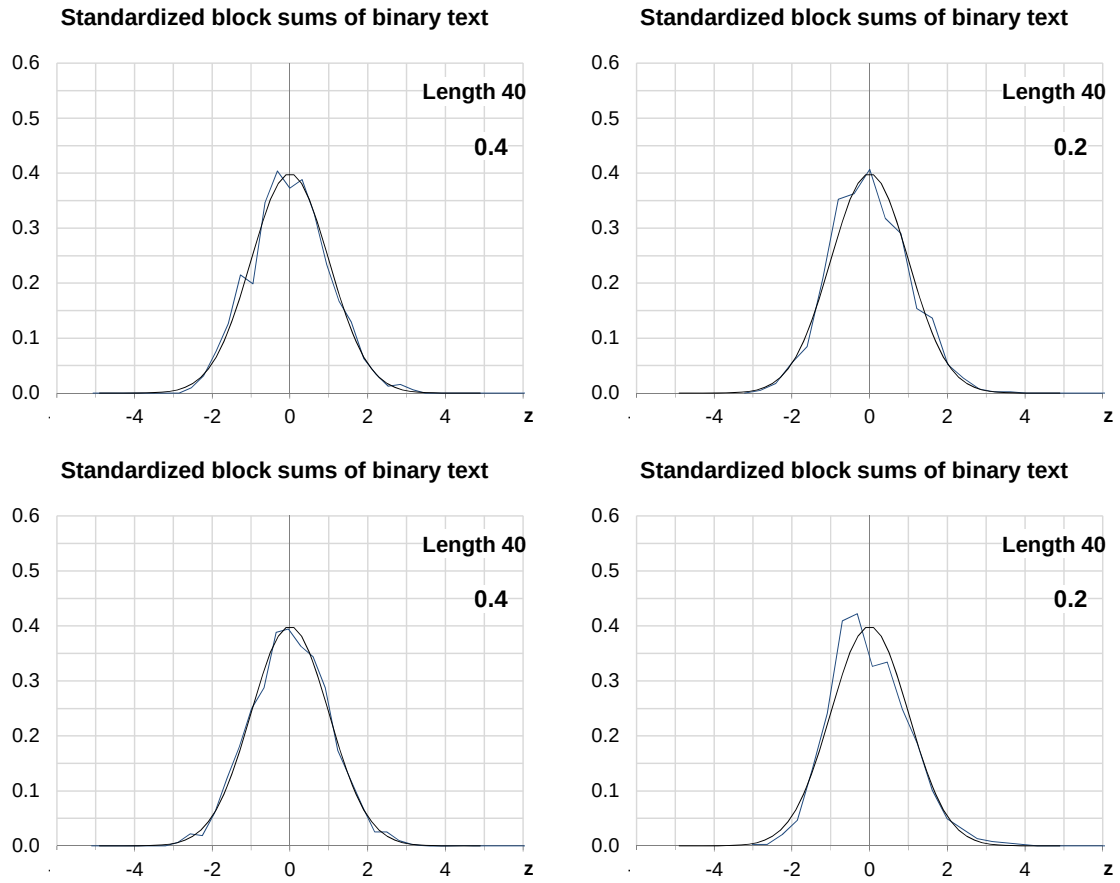

Figure 10. Two replications of binary text – distribution of standardized block sums; Left: fairly symmetric with p =0.4, Right: skewed to the right (steeper on left side) with p = 0.2

3.2. Describing the generation of text blocks by the binomial distribution

Remark on simulation: The variation of binary data (0, 1) for a random sample of size 1,000 is roughly 0.03, i.e., a probability can be estimated with that precision but not more precisely (if we allow for a "risk" of 5%). That means, our 1,000 data should not be over interpreted as additionally there is this source of random variation. Deviations in the simulation scenario can be caused by the low precision of simulation or by bad fit of the standard normal curve. We will eliminate the effect of simulation by using the binomial distribution, which applies for our random generation of binary text. The new method will let us see the increase in fit from increasing the length of text much better and free of the "noise" of simulation.

If the text is generated by random numbers that attain the value of 1 with probability p and 0 with $1-p$ (and the random numbers behave as if they are independent) then the single signs of the first text block of length 20 are random variables $X_{1,1}, X_{1,1}, ..., X_{1,20}$ (first index for the block number and second for the number of the sign within the block) and the block sum $B_1 = X_{1,1} + X_{1,1} + ... + X_{1,20}$ follows a binomial distribution with n = 20 and p. Rather than continue with simulating the data for the other blocks we will describe the potential outcome by probabilities from this binomial distribution. The probabilities can be interpreted as idealized frequencies. If we describe the situation in block i then we have an analogue situation: the block

sum $B_i = X_{i,1} + X_{i,1} + ... + X_{i,20}$ follows the same binomial distribution. Mean and standard deviation of the block sum can be estimated from the data of all 1,000 blocks or they can be predicted from the mean μ and standard σ deviation of the binomial distribution, which are: $\mu = n \cdot p$ and $\sigma = \sqrt{n \cdot p \cdot (1-p)}$.

The frequency polygon described the distribution of the standardized block sums; these data are generated by the standardized random variables $\frac{B - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$ (we have omitted the index for the block number). The close fit of the standard normal curve means also that we can approximate the distribution of *B* (a binomial distribution) by the normal distribution. More precisely, we can approximate: $B(n, p) \approx N(\mu = n \cdot p, \sigma = \sqrt{n \cdot p \cdot (1-p)})$. What we also have found is that the fit is better for $p = 0.4$ than it is for 0.2. We will now compare various binomial distributions with the corresponding normal distribution. We will no longer standardize the values but remain in the original scale of the block sum.

3.3. Various diagrams to display a discrete distribution

Several graphs for a discrete distribution are compared to each other. All have their relative merits. We will use the shadow graph as it supports an area representation, which becomes relevant if a discrete distribution is compared to a continuous distribution.

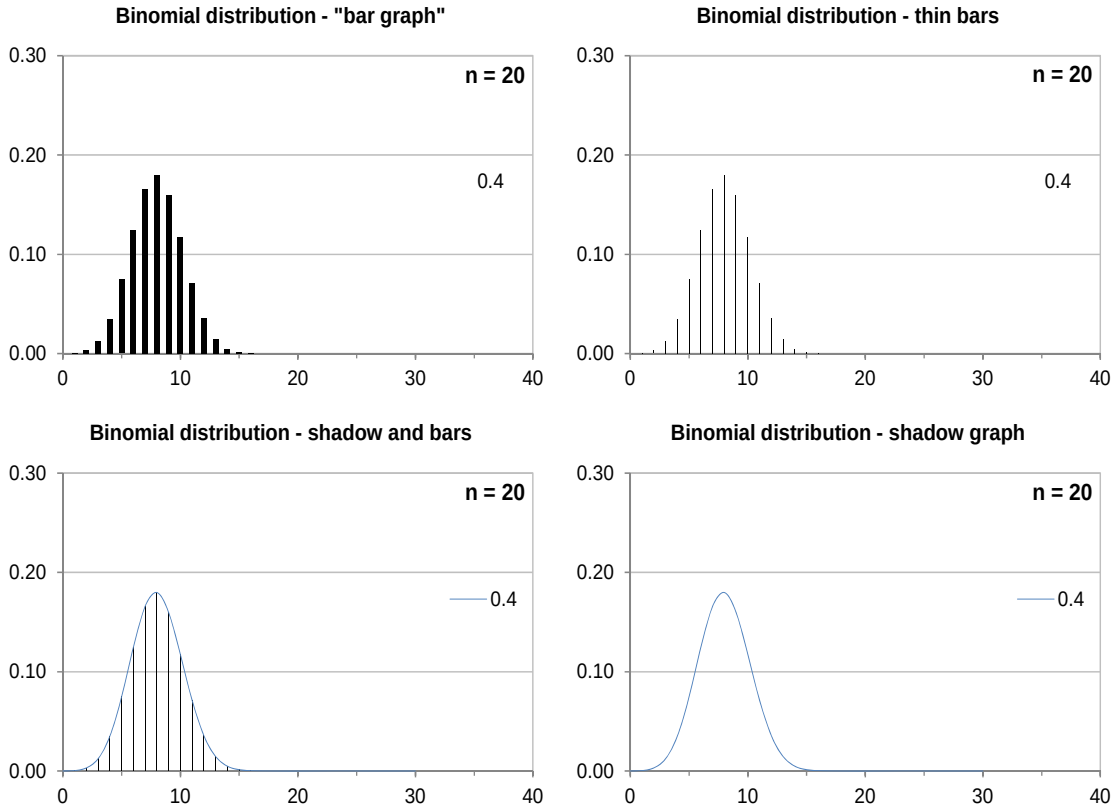

Figure 11. Bars, thin bars, and shadow graph to represent a binomial distribution

First we will introduce a shadow graph (a probability polygon) for the binomial distribution. Usually (as in Figure 11), the binomial distribution is illustrated by a thick bar graph though only the single points 0, 1, …, n have probabilities distinct from zero. Thus, a thin bar should represent the probability. However, in comparing the discrete binomial to the continuous normal distribution, the area becomes the key to convey the probabilities. Thus, we replace the thin bars by a shadow line connecting the top of the bars. Usually we remove the thin bars from the graph. It is these shadow graphs we compare to the normal curves (not the standard normal curve but those in the original scale of the sums).

3.4. Analysis of artificial binary text by inspecting binomial distributions

In the following, we will not generate more text and analyse it. Instead we will use our knowledge about the binomial distribution that describes the probability distribution of the block sum. Rather than basing our analysis on the distribution of standardized block sums, we will work with the original values of our block sums and present their distributions by the shadow diagram and additionally vary the probability p for the sign 1. In a first step we will see the shape of the distribution (left column of Figure 12), which is fairly symmetric for middle values of p. In a second step we will draw the normal distribution for comparison (right column of Figure 12). The shadow graphs look nearly like normal curves though there is a definite skewness for $p = 0.1$ and 0.9. If we draw the corresponding normal curve for comparison, we can see the good fit. The comparison to the normal curve makes the skewness even more apparent. We repeat the comparison with block length of $n = 40$. According to the usual recommendations, the normal approximation is not yet allowed as the rule of thumbs requires that $n \cdot p \cdot (1-p) > 9$, which is only fulfilled in our best case in the middle line of Figure 13. However, for $n = 100$ all cases fulfil the 'rule' and the graphs show a good fit (Figure 14).

Here, the improvement of the normal fit by the increase from a block length of 20 to 40 just gives a qualitative impression that the fit should improve in the sense of a mathematical limit theorem, the Central Limit Theorem. If the block length n is increased to infinity, the normal curve should be the limiting function. It seems clear that such a theorem – for mathematical reasons – has to refer to *standardized* block sums instead of block sums. For the binary text generation, the block sums tend to have a mean of $n \cdot p$ and a standard deviation of $\sqrt{n \cdot p \cdot (1-p)}$, which both increase beyond any constraint so that the block sum has no distribution at all in the limit. It is continuously shifted to the right and gets flatter till no distribution remains. That is the reason for the initially awkward standardization that has been introduced in analysing the text. A final sequence of graphs for $n = 100$ should convince the reader that such a limit theorem should hold (Figure 14). We cannot see a difference between the binomial shadow and the approximating normal curve even in the worst case of $p = 0.1$.

3.5. Heads minus Tails – analysis of a game instead of texts

We play coin tossing games and investigate the balance of number of Heads minus number of Tails. We produce our "text" now by an experiment that has only two signs, 1 (Heads) and –1 (Tails). We introduce again blocks, i.e., we combine 20 signs to form one block, and calculate the block sum, which is the balance of a player who bets on Heads against a casino if the player wins 1 Euro or loses 1 Euro depending on the result of the toss. The block sum represents the balance after the games of a block have been played. We are interested in the distribution of the player's balance for playing one block. To find this distribution we can simulate the game or we can use the binomial distribution with n as the block length and $p = 0.5$ for an ideal coin.

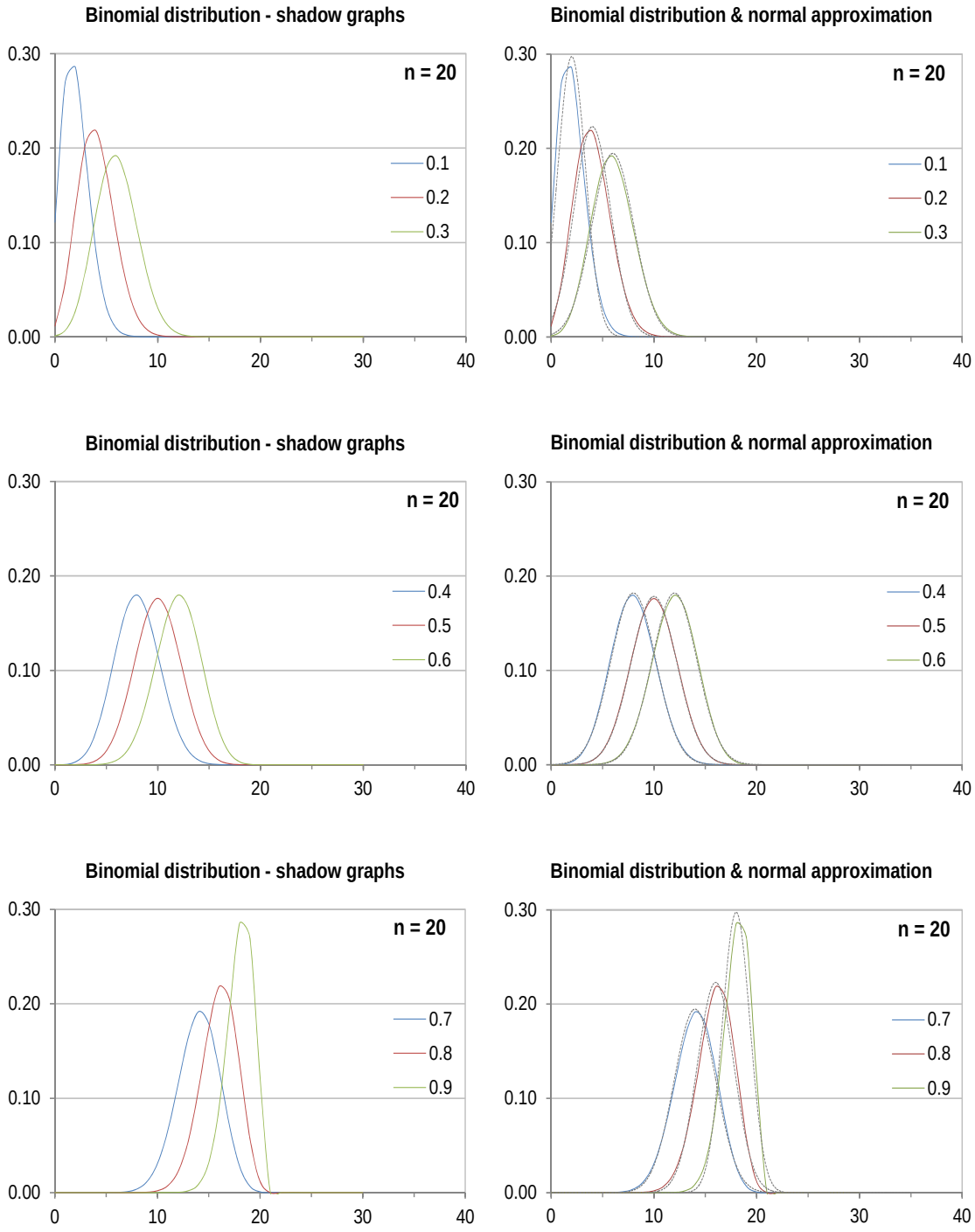

Figure 12. Inspection of the shape of binomial distributions with $n = 20$ and comparison to the normal distribution (dashed curves) in the right column of the figure

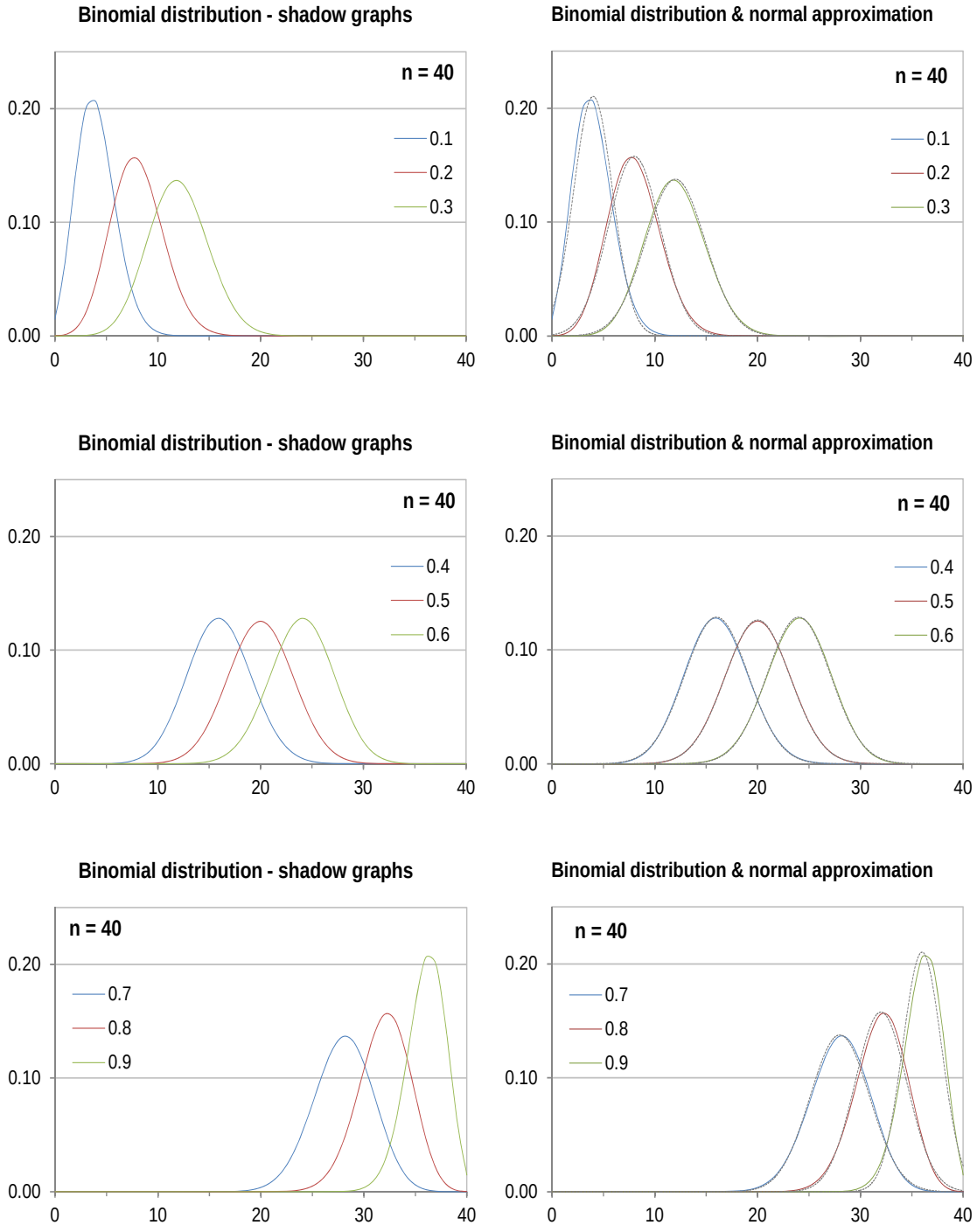

Figure 13. Inspection of the shape of binomial distributions with $n = 40$ and comparison to the normal distribution (dashed curves) in the right column of the figure

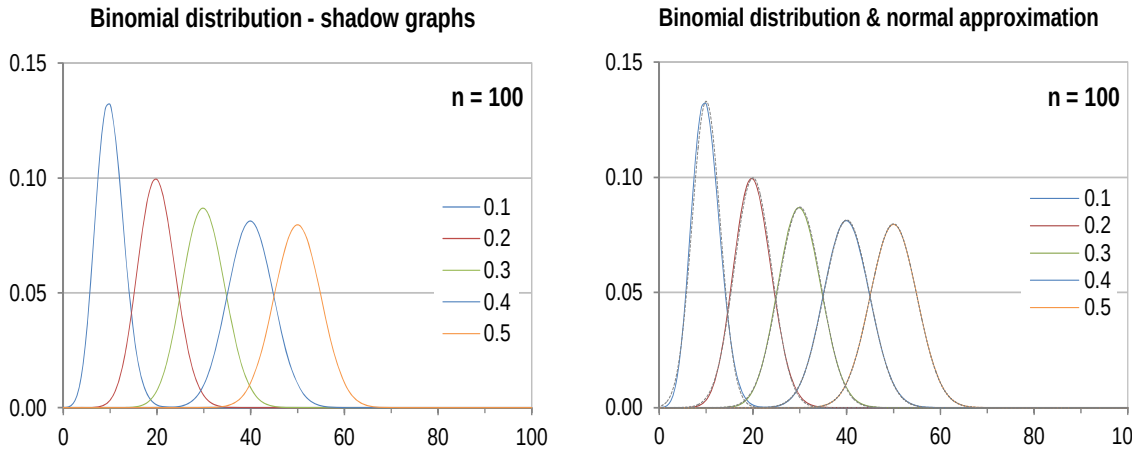

Figure 14. Inspection of the fit for block length 100 for values of p from 0.1 to 0.5

The number of values increase from 41 (for $n = 20$ trials) to 81 and finally to 201 (for $n = 100$). The spread increases: visible bars (probability greater than roughly 0.05) from −12 to 12 (for $n = 20$ trials) to −18 to 18 and finally to −24 to 24. The single result of an experiment is the sum of n tosses (1 for head and −1 for tails) so that we expect that the standardized values of Heads minus Tails will approximately follow a normal distribution. The systematic error of a continuous distribution that should replace the discrete bars gets smaller and smaller as with the standardization the gaps get smaller.

The whole range of the random variable Heads minus Tails *is rescaled* to roughly −5 to 5 (or even to −4 to 4). For n increasing, the Central Limit Theorem states that finally (a thought experiment, which can never be reached in real world as we cannot perform an experiment an infinite amount of times) the distribution of *the standardized* variable Heads minus Tails reaches the standard normal curve. This limiting statement (a mathematical theorem) gives a justification to approximate the distribution of Heads minus Tails (on the original scale) by a normal distribution. The original scale is regained from standardized values simply by a linear transformation, i.e., a scaling and a shift, which both preserve the shape of a normal distribution (only mean and standard deviation change from 0 and 1 to shift parameter and scaling factor).

In Figure 15, we see that the distribution of Heads minus Tails remains centred around zero. However, its spread is increasing without limit. There is no limiting distribution for Heads minus Tails. The limiting distribution occurs only for the standardized variable Heads minus Tails, i.e., we have to subtract 0 and divide by its standard deviation.

This is how the Central Limit Theorem helps us to approximate FINITE SUMS of the inspected random variables (the sum of the single tosses encoded with 1 and −1 here). As the AVERAGE (mean value) of the data, i.e., the SUM divided by the number of trials is also a rescaling, we get a justification to approximate the distribution of the average of random variables by a normal distribution. For practical reasons we are neither interested in standardized sums, nor in sums but we focus on the average (mean value) of random variables as this will help us to estimate the mean of the population from which the single variables pick out one element randomly. This population is often thought to be finite but in mathematics we can also think of the population as a *process*: the process of tossing a coin, e.g., which is a random variable. And in our present setting, this random variable attains the value of 1 if Head occurs, and −1 if Tail occurs.

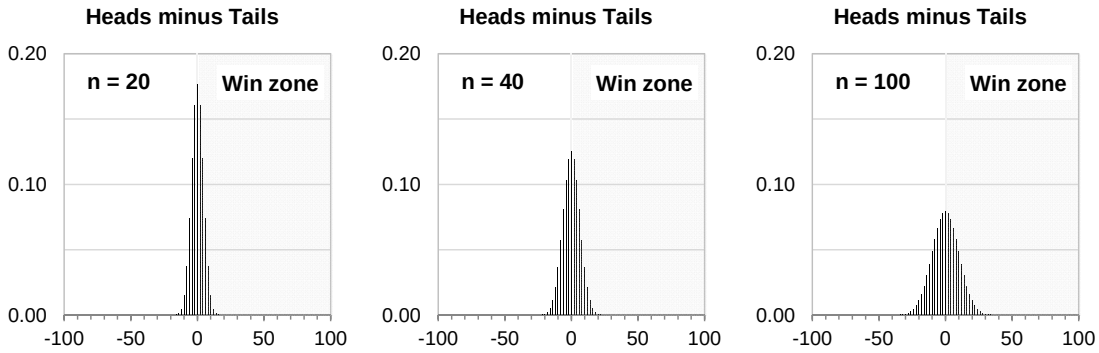

Figure 15. Balance of Heads minus Tails for a fair coin after *n* trials

An interesting consequence is seen from variants of the coin tossing game. If a biased coin is used ($p = 0.4$) then the player can have a positive balance after 20 or even 40 trials but we see that chances are getting much smaller with 100 trials (Figure 16). The risk (the probability) for high losses has increased substantially after 100 trials. These properties get more distinct if the number of trials is increased. In the long run, the casino will surely win. Of course, the casinos will usually offer a less biased game when the player can win for a longer time but finally will also lose all money. The chances for simple bets on the roulette table are 0.4865 (18/37), for example. That also keeps players to continue their games as they think they have their own systems to beat the casino (see Figure 17).

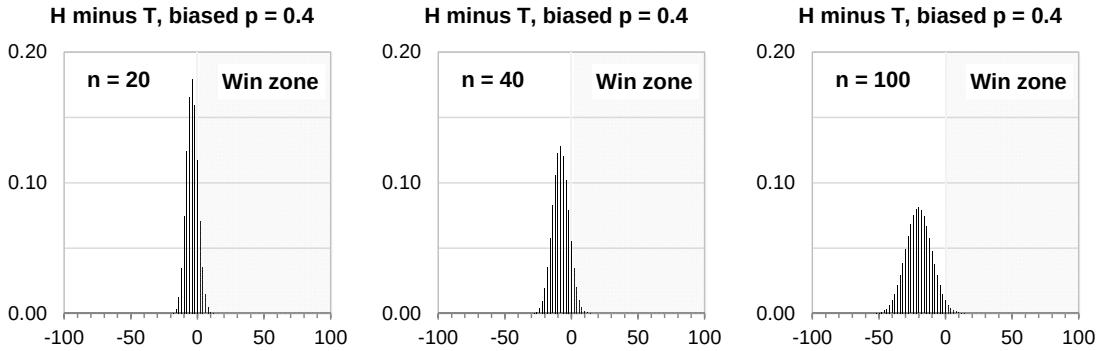

Figure 16. Balance of payments for a biased coin ($p = 0.4$)

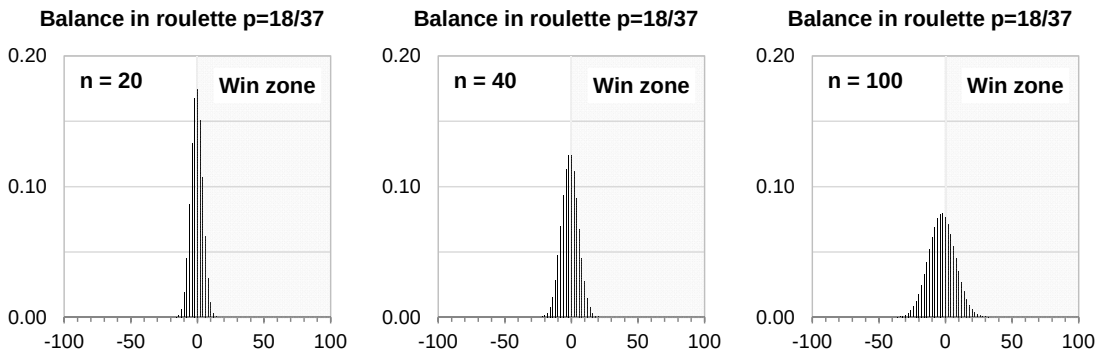

Figure 17. Balance of payments for roulette betting on pair / impair or rouge / noir

We did introduce the game of Heads minus Tails not only to illustrate the bad perspectives of players in the casino. We introduced it also as a special case where the Central Limit Theorem may be proved by relatively easy mathematical tools that are within the reach of brighter secondary students. It may be important to give at least a mathematical argument why such a theorem should hold. The simulation studies yield only restricted empirical evidence for such a theorem and can clarify circumstances under which such a law can hold. However, the simulation per se cannot replace a proof and it leads also to confusion as – obviously – we cannot continue experiments infinitely many times in real world. The way to prove the special case of the Central Limit Theorem follows closely the path of de Moivre when he first introduced the expression of the normal density in approximating the binomial probabilities for the Heads minus Tails distribution. He investigated the absolute values of this random variable applying Stirling's formula for *n* factorials to approximate the harmonic series involved in the proof.

4. Describing the original task more formally

For the original task of the text with the full set of signs and the block sums we can reformulate the situation and the calculations now analogously to our considerations with the binomial distribution. In each block, the sum is thought to be generated by more general random variables than the wheel of fortune with only two sectors 0 and 1. We can think of a wheel with sectors that correspond to each sign used in the text with an area that corresponds to the frequency of the sign used in the text (Figure 18).

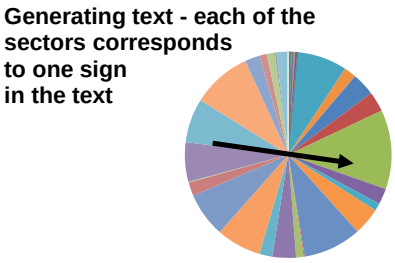

Figure 18. Generating a block of length 20 by spinning the wheel 20 times

4.1. Block sums as random variables and their distribution

The block sum is represented by adding the result of 20 times spinning the wheel, which leads to the random variable $B_{i,20} = X_{i,1} + X_{i,2} + ... + X_{i,20}$ now with a general wheel as in Figure 18. Again, the close fit we have found for the standardized block sums is expressed by the mean and standard deviation of the wheel; note that we have attributed numerical values (codes) to the signs. The approximate distribution for any standardized block sum (we have found this relation for our frequency polygon on the data for block length 20) is standard normal:

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

We could likewise state that $B_{20} \approx N(\mu_{20}, \sigma_{20})$ by rescaling our standardized data back to the original scale. It remains to confirm that the mean and standard deviation for a block of length 20 are related to the mean and standard deviation of the wheel (that describes how a

single sign is produced) by: $\mu_{20} = 20 \cdot \mu$ and $\sigma_{20} = \sqrt{20} \cdot \sigma$ (we can give theoretical reasons or check our data whether such a relation should hold).

4.2. Central Limit Theorem (CLT)

The Central Limit Theorem can now be formulated as: We have *n* independent random variables $X_1, X_2, ..., X_n$ that all have the same distribution as X, which has a finite expected value μ and a finite standard deviation σ. We define the sum as $B_n = X_1 + X_2 + ... + X_n$ with an expected value μ_n and a standard deviation σ_n. The standardized random variable $\widetilde{B}_n$ is obtained by $\widetilde{B}_n = \frac{B_n - \mu_n}{\sigma_n}$. Its cumulative distribution function is $F_n(z) = P(\widetilde{B}_n \le z)$; the cumulative distribution function of the standard normal distribution (with expected value 0 and standard deviation 1) is denoted by Φ. Under these conditions, the following limit theorem holds:

$$\lim_{n \to \infty} F_n(z) = \Phi(z) .$$

We will read this back in terms of our text "analysis". *X* is the generic term for the generation of a sign in the text and may be thought of as a special wheel of fortune. X_2, e.g., is the second spin and describes how the second sign is generated and a numerical value (like the ASCII code) is assigned to it. We generate *n* signs for one block of length *n*. The different spinnings of the wheel are intuitively thought as independent trials, which correspond to the mathematical independence of the random variables $X_1, X_2, ..., X_n$. We have noted that in natural texts, this independence is violated and we have tried to introduce independence between signs within a block by a random rearrangement. The random variable B_n describes how the block sum is made up of the values that correspond to the single signs. From the data b_n of many (1,000) blocks, we estimated the mean and the standard deviation of the block sum: $\mu_n \approx \overline{x}_{b_n}$ and $\sigma_n \approx s_{b_n}$. Thereupon we built the standardized block sums $\widetilde{b}_n = \frac{b_n - \overline{x}_{b_n}}{s_{b_n}}$, which are data for the standardized random block sum $\widetilde{B}_n = \frac{B_n - \mu_n}{\sigma_n}$. We inspected the distribution of the standardized block sums and found a good fit of the standard normal curve. If the block length *n* is increased beyond limits (which is only a thought experiment), then the investigated distribution approaches the standard normal curve. This is the statement of the Central Limit Theorem (CLT) in its simplest form (LeCam, 1986, describes the thrilling history of this theorem and its generalizations that have brought forward the need for axiomatizing probability).

We have inspected the frequency polygons for *n* =20, 40, and 100 (the latter only for the artificial text generation) and found out that they come close to the normal curve; we could also investigate histograms with the same result. Thus, our investigation establishes empirical evidence for the CLT. From the CLT we derive a justification to approximate the distribution of the standardized block sum $\widetilde{B}_n$ for *finite* n. If it converges then we should be close to the limit after *n* is large enough. The question remains, how large *n* should be and we have found sufficient precision already with values of *n* = 20. The debate of the size of *n* depends mainly on the kind of distribution we use for the wheel that describes the generation of a single sign (or better, the distribution of the values that are attributed to this sign). Block lengths of *n* = 100 were not sufficient for *p* = 0.1 for our artificial binary text generation (though the resulting distribution

was only slightly skewed). For smaller values of *p* even longer block lengths are needed to attain a reasonable fit for the standardized block sums.

4.3. Implications of the CLT – normal approximation of sums and averages

Once we have established reasons for approximating the distribution of the standardized block sums by a standard normal curve, we can also use these reasons for approximating the block sums on the original scale by a normal distribution. We only have to adapt the parameters from 0 and 1 to the shift parameter and the scaling factor that have been used to standardize the block sum, i.e., the fitting normal distribution has a mean of $\mu_n = n \cdot \mu$ and $\sigma_n = \sqrt{n} \cdot \sigma$. Analogously, the mean value of a block $M_n = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{B_n}{n}$ establishes only a further rescaling of the standardized block sums and therefore its distribution can be approximated by a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. We will need this result for statistical inference later. The relations of mean and standard deviation for the various statistics of a block can be estimated empirically from our data. In the generation of binary texts we can also use our knowledge about the mean and standard deviation of the binomial distributions, which comply with the equations above. We could also give intuitive arguments for special cases of random variables that support the given relations (see Borovcnik, 2001 and 2011). A general proof, however, requires more mathematics. Thus, we might prefer to support the properties by analysing data from computer simulations.

Note: There is neither a Central Limit Theorem for the sums, nor for average values of blocks. Both random variables have no limiting distribution (see Figure 19). While the sums tend to get larger and flatter, till no distribution remains, the averages remain centred but their spread gets smaller till the value coincides with the centre axis (at the expected value of the wheel that generates single data), which corresponds to a generalization of the Law of Large Numbers. However, as we have a justification for the normal approximation of the standardized block sums, we can use this argument as both block sum and block average are linearly rescaled from the standardized sums. Rescaling does not affect the fact that a normal distribution applies as approximation. Yet, of course, it influences centre and spread of the fitted normal curve.

5. Samples and populations – statistical inference

We have analysed so far a factual text or a generation method to produce binary text. For that reason we have split the (generated) text into blocks and investigated the distribution of the block sum. We will look on the analysis with a statistical eye and regard the text as population and the text blocks as samples from which we want to extract information on the text.

5.1. Reinterpretation of text analysis in terms of samples and populations

We will re-interpret the text as the population to be investigated. If the text is finite we speak of finite populations, if the process to produce text (potentially infinite text) is investigated, we will speak of infinite populations. The distribution of the ASCII codes in the whole text turns to the parent distribution, from which we take our samples. Likewise, the generation method to generate binary symbols by a wheel of chance will be called the parent distribution. As we have mentioned, we can also use more complicated wheels of chance to generate text.

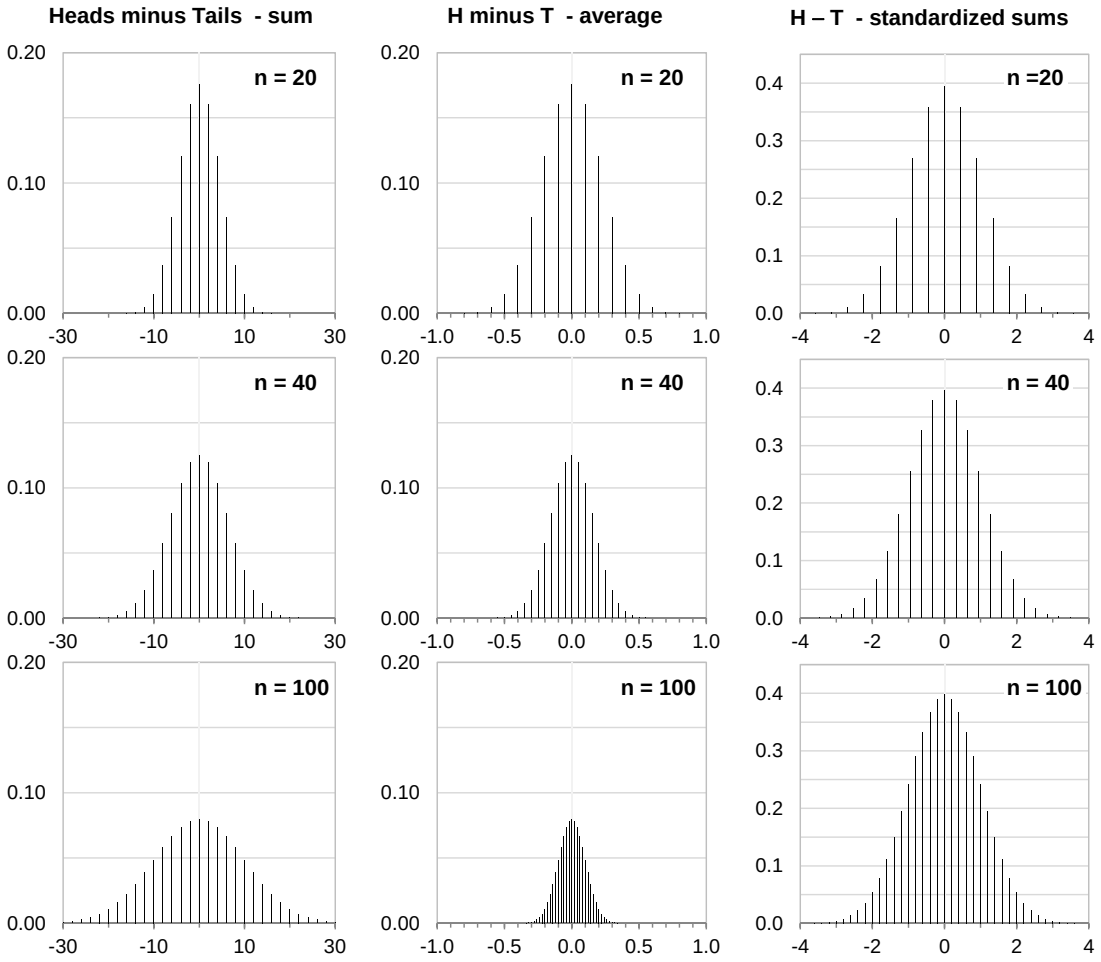

Figure 19. The sum diverges (left column) – the average converges to a single point (middle) – the standardized sum converges in distribution to the standard normal curve (right column)

The text blocks, which we have analysed, become the samples in this view. Our random rearrangement of texts improved the fit of the standard normal curve for the distribution of the standardized block sums. That highlights that we should have random samples as our text blocks. Random blocks or samples guarantee that single signs can be exchanged without changing the general feature of the text blocks. In inferential statistics we introduce methods of generalizing information on the population (the parent distribution) from the data of a sample. In our text analysis, we could be interested in the mean value of the population (all ASCII codes with their frequency, or the wheel with $p = 0.4$ in the binary text generation) from the information of the text blocks. In Figure 20, the distribution of the ASCII codes in the text (left) or the bar graph (right) yields a static view of the text, how the various values attributed to the signs are distributed. The wheel represents a dynamic view on the same population; by spinning it, text blocks (samples) can be generated, which purport the information on the population. The process of generating text becomes the population, while the generated text blocks turn to samples.

In general, we have only *one* text block, i.e., one sample of length *n*. In order to investigate the relation between the block sum divided by the length of the block (average value of the signs in the sample) and the average in the population we use the distribution of the characteristic under scrutiny; i.e., how does the block sum (the average, etc.) vary if another sample

(block) is analysed? We have investigated the distribution for the block sum (the sum of values for a sample). The Central Limit Theorem states that the standardized block sum may be fitted by a standard normal curve. This gives a justification to approximate the distribution of the block sum by a rescaled normal distribution. Furthermore, it gives a justification to approximate the distribution of block (sample) averages by a normal distribution.

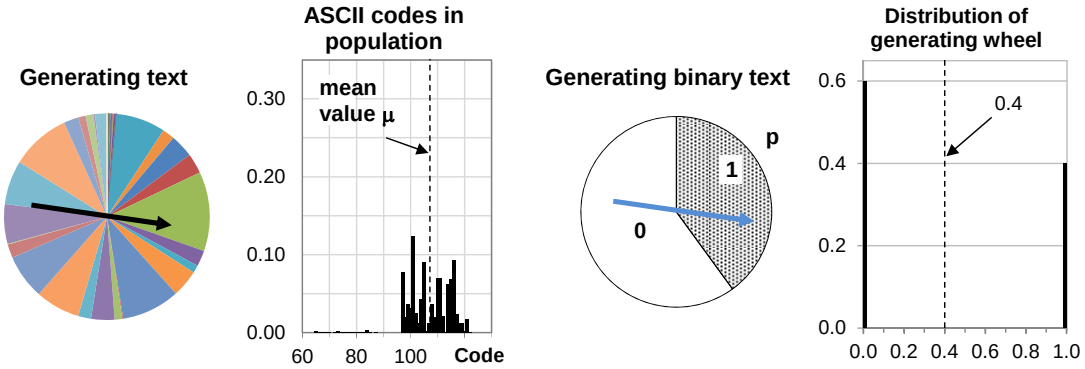

Figure 20. Representations of population and parent distribution–Generating wheel & bar graph

5.2. Estimating a mean of a population from blocks or samples

Statistical inference has to do with studying the interrelations between the generating wheel and the produced text. We show how the new view of population and samples gives a justification to conclude from the single value (one average of the data) in a sample backwards to the average of the population. The shrinking of the distribution around the value of the population mean corresponds to the Law of Large Numbers (for means). The Central Limit Theorem will provide numerical probabilities that certain specified threshold values will be violated by the average in a sample of specific size *n*. We show two series of figures (Figure 21 and 22), one for general wheels (for general "text") and one for our procedure to generate binary texts. The method to produce text becomes the method to draw a random sample in the new setting.

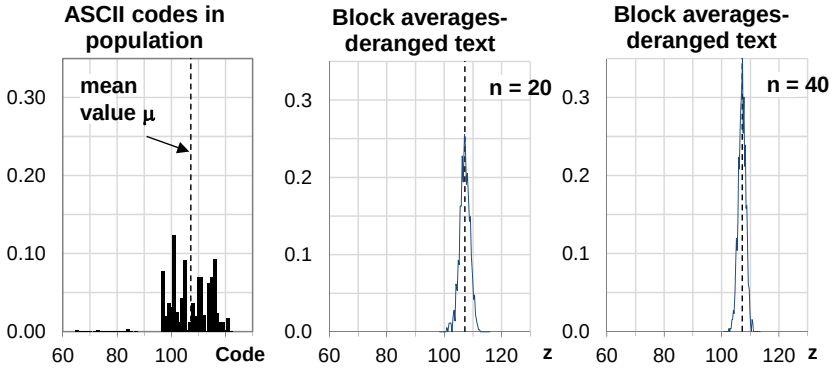

Figure 21. Distribution of ASCII codes in the population that generates text (as represented by the wheel in the previous figure and the distribution of the average value of blocks/ samples.

Figure 21 shows that the mean of the population (which is also the mean of the wheel that generates text) is reflected in the averages of text blocks: mean values of single blocks are scattered around the same "axis" signified by the mean of the population. The averages of text blocks lie closer to this axis if the block length increases. As a thought experiment – they will restrict to a single point (the axis) if the length is increased without limit.

5.3. Estimating a proportion from the "average" of blocks or samples

For binary text, the relations between the generating wheel and the text blocks 'produced' are analogue to the general wheel. From Figure 22, we can see how the distribution of block (sample) averages restricts to the axis, which is determined by the mean value of the population (wheel). Again by an idealized experiment, we can imagine how the distribution shrinks to this axis if we investigate longer text blocks (larger samples, ideally unlimited).

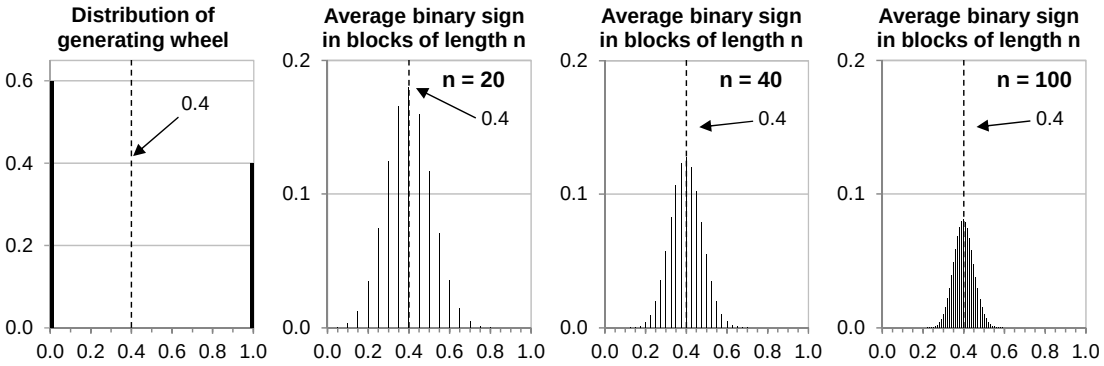

Fig. 22. Binary text of varying length – the average from samples reflects the probability for 1's

5.4. Implications of CLT and LLN for sampling

The Central Limit Theorem guarantees that deviations beyond a threshold can be calculated by the normal distribution, not only for the sample sum (block sum) and sample averages (block averages). In generalizing the result of the CLT, we can approximate the distribution of any statistics that is defined by a sum of values of single elements of the sample by the normal distribution. For example, the distribution of the sample variance (the square of the standard deviation) follows a chi-square distribution. However, from the CLT we will expect that – for larger samples – this will come close to the normal distribution. The considerations and experiments in the present paper explain why the normal distribution has become so important for inferential statistics. For the Law of Large Numbers there are some nice simulations and experiments in Borovcnik (2001) or in Borovcnik and Schenk (2012).

References

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

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 (Simulations in teaching stochastics). *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG)*, 44, 1-16.

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.