DIMENZIÓK
Matematikai Közlemények (Mathematical Notes)
XIII. kötet, 2025 13–24

doi:10.20312/dim.2025.02

Computer-aided validation of basic concepts of probability theory for high school students

Sándor Zsuppán
Berzsenyi Dániel Evangélikus (Líceum) Gimnázium és Kollégium
Sopron, Hungary
zsuppans@gmail.com, 0009-0006-3454-1528

ÖSSZEFOGLALÓ. A középiskolai matematika tananyag részét képező valószínűség, feltételes valószínűség, valószínűségi eloszlás és várható érték fogalmak számítógépes segítséggel történő szemléltetését mutatjuk be. Az internetről hozzáférhető véletlenszerűen vagy pszeudorandom módon generált, viszonylag sok adatot tartalmazó adathalmazokkal dolgozunk, hogy a nagy számok törvénye alapján a fenti fogalmakat a relatív gyakoriság és átlag fogalmakkal állíthassuk párhuzamba. Kitekintésképpen a Bernstein-polinomokhoz hasonló, de a hipergeometrikus eloszlásból származtatott polinomokat vizsgálunk röviden.

ABSTRACT. In this note we cover some examples of computer-aided illustration of the concepts of probability, conditional probability, probability distribution and expected value for high school students. We utilize relatively big true random or pseudorandom datasets available on the internet or generated locally in order to be able to make use of the law of large numbers. Apropos of an example we briefly investigate Bernstein like polynomials derived from the hypergeometric rather than from the binomial distribution.

1 Introduction

Several basic concepts of statistics and probability theory are contained in the requirements [1] for high school graduation in mathematics. These include at the intermediate level exam e.g. relative frequency and average in statistics and their parallel concepts probability and expected value in probability theory, respectively. Ref. [1] requires at the advanced level additionally the concepts of conditional probability and probability distribution, especially the binomial and the hypergeometric ones for random sampling with or without replacement, respectively. Even though the law of large numbers in its precise form is not directly specified in [1], the students should use it intuitively by knowing that although the outcome of a random event is not predictable, the relative frequency of the favorable outcomes from many mutually independent random samplings is a good approximation for its probability. Moreover, in the same sense the average of many mutually independent random samplings approximates the expected value. These correspondences facilitate the utilization of randomly generated big datasets and of the

HUNGARIAN TITLE. Valószínűségszámítási fogalmak számítógépes illusztálása középiskolásoknak
KULCSSZAVAK. valószínűség, számítógéppel támogatott, Bernstein polinom.
KEYWORDS. probability, computer-aided, Bernstein polynom.

computers processing them for demonstrating the above mentioned concepts. On the one hand such datasets in various subjects are easily and mostly freely available on the internet, on the other hand basic programming skills are also required for high-school graduation in informatics not only on the advanced level but also on the intermediate level exam. Hence experimenting with the processing of such big datasets on a computer can improve not only the mathematical but also the programming skills of the students. A similar concept for computer-aided illustration of interesting mathematical problems for high school students was treated in [6].

In this note we discuss some examples, wherein the used datasets consist of true random numbers from the internet or pseudorandom numbers generated locally by the computer. As recommended for the graduation exam in informatics we utilize the Python programming language (python.org) along with the useful libraries NumPy (numpy.org) and SciPy (scipy.org) for calculations and Matplotlib (matplotlib.org) for graphical visualization in the Jupyter Notebook interface (jupyter.org). This note is not intended as a full description of the teaching methodology of the subject, we just demonstrate some interesting exercises in the first section. In one of these exercises some polynomials pop up, which resemble the Bernstein polynomials [2, 5], however, these polynomials emerge from the hypergeometric distribution rather than from the binomial one. In the final section we briefly investigate some basic properties of these Bernstein-like polynomials.

2 Exercises for probability theory with big datasets

In this section we discuss some interesting probability theory exercises. The results will be visualized using relatively big datasets stored in comma-separated value (csv) files which Python can handle.

Exercise 1. Simulate many rolls of a 6-sided fair die and plot the variation of the relative frequencies of all possible outcomes as functions of the number of rolls! Compare the diagram to the probabilities of the respective outcomes! Plot also the variation of the average of the outcomes as a function of the number of rolls! Compare the diagram to the expected value of the outcomes!

The probability of each outcome of a die roll is $\frac{1}{6}$ and the expected value equals 3.5. We use the 'Random Integer Generator' item on random.org and save the generated random integers between 1 and 6 (both inclusive) as a 'csv' file. (This service is free up to 10000 random integers.) We can observe that according to the law of large numbers the relative frequencies and averages of the outcomes tend (however oscillating) toward but at the same time oscillate around the probability $\frac{1}{6}$ and expected value 3.5, respectively.

Exercise 2. Generate many random digits independently and let move a random walk along the number line as follows: start at zero and move a distance which equals the random digit left if the digit is even but right if it is odd. Let the random variable X_n denote the position of the random walk after n steps! Calculate the expected value and the standard deviation of X_n! Plot the position of the generated random walk as a function of the number of steps and compare the outcome to the expected value and to the standard deviation! Treat the digits of some irrational numbers as if they were generated at random, plot the connected random walk and compare it to the results of the previous two parts!

If the digits are selected with equal probability $\frac{1}{10}$, then the expected value of one step of the walk equals $E_1 = \sum_{k=0}^{9}(-1)^{k+1}k = \frac{1}{2}$ and for the standard deviation we have $D_1^2 =$

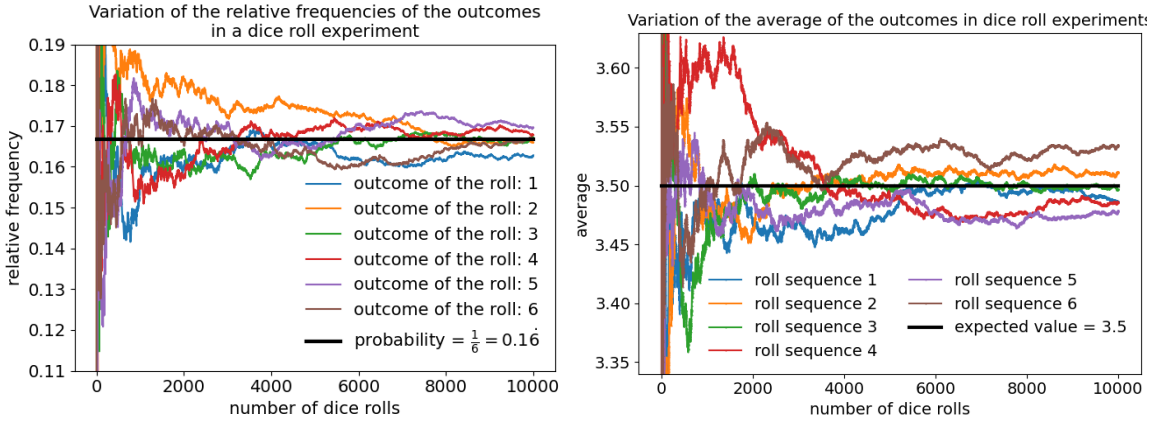

Figure 1. Simulating dice rolls

$\sum_{k=0}^{9}\left((-1)^{k+1}k-\frac{1}{2}\right)^2 = 282.5$ and hence $D_1 = \sqrt{282.5} \approx 16.8$. Because the steps in this random walk are mutually independent there follows $E(X_n) = \frac{1}{2}n$ and $D^2(X_n) = nD_1^2$ which implies $D(X_n) = \sqrt{282.5n} \approx 16.8\sqrt{n}$. According to the diagram on the right of Figure 2 the

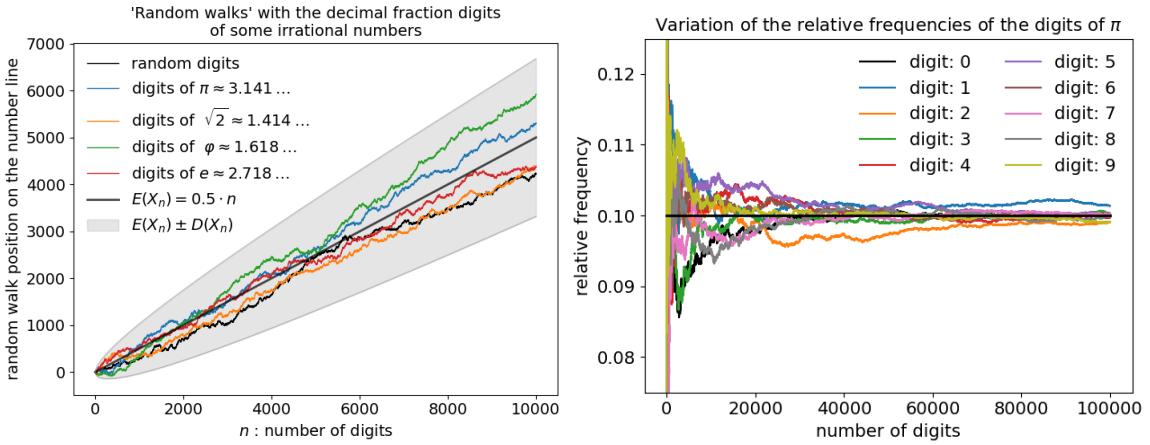

Figure 2. 'Random walks' with irrational numbers

relative frequency variations of the decimal digits of π pretend to be randomly generated with discrete uniform distribution, however this is not yet proven, see e.g. [4] and the references there. On the left diagram of Figure 2 we can see that using the irrational numbers π, $\sqrt{2}$, the golden ratio φ and the Euler constant e as 'random digit generators' give very similar results to an actual random walk with digits generated on random.org. (The digits of the used irrationals can be found in e.g. https://apod.nasa.gov/htmltest/rjn_dig.html and https://www2.cs.arizona.edu/icon/oddsends/phi.htm.)

Exercise 3. Let the random variable R denote the distance of a random point in the unit square $[0;1]^2$ from the origin. Calculate the probability $\mathrm{P}(R \leq 1)$! Determine the distribution function of the random variable R and compute its expected value! Place many random points into the unit square, calculate the relative frequency of those with $R \leq 1$ and compare it to the probability! Compute also the average distance of the points from the origin and compare it to the expected value!

In order to produce the diagrams on Figure 3 we used the 'Decimal Fraction Generator' item on random.org to obtain two times 10000 random decimal fractions in the interval [0;1]

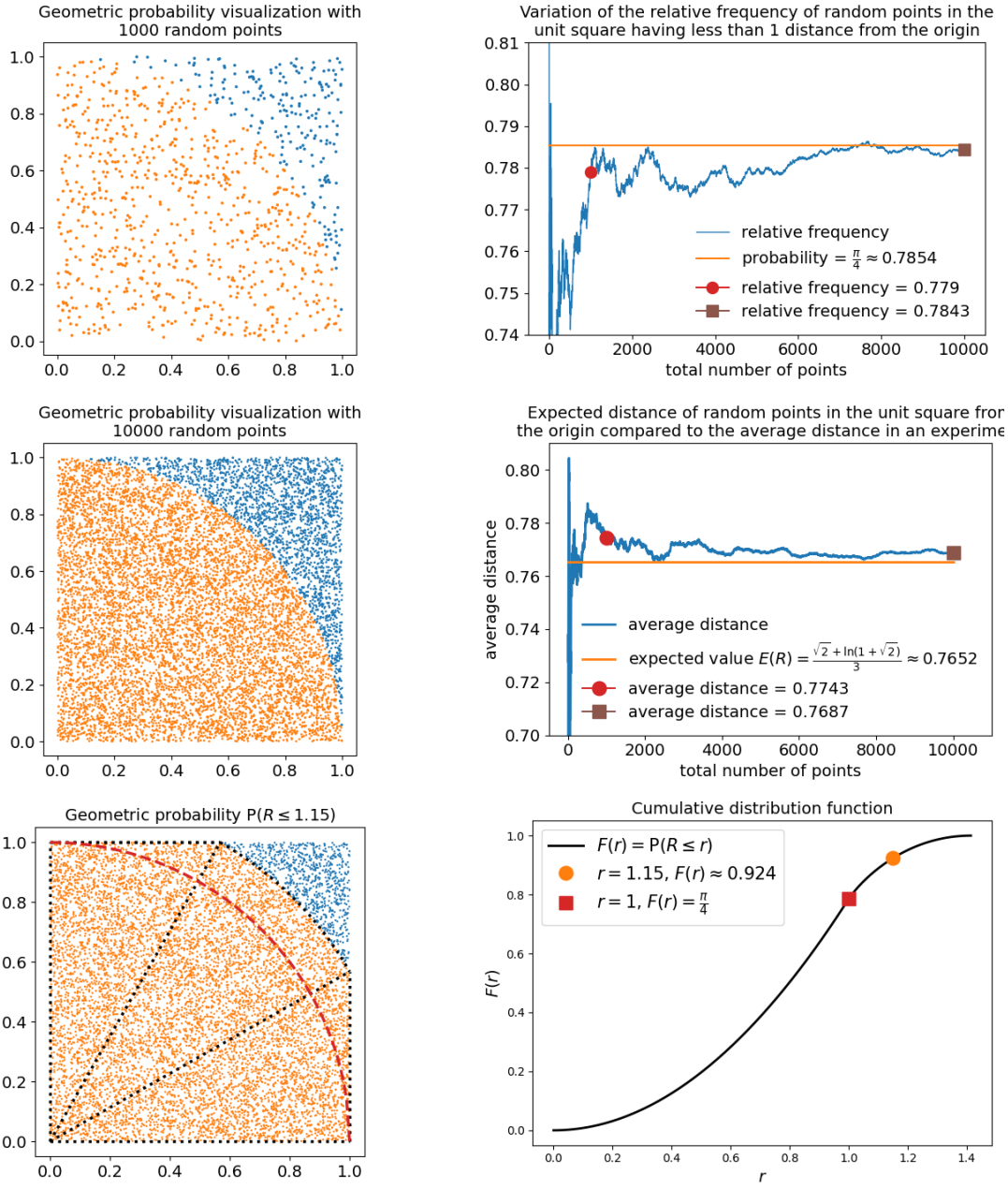

Figure 3. Geometric probability exercise

which then ordered in pairs serve as coordinates for points in the unit square. The exercise can be solved with geometric probability calculating the corresponding areas (depicted on Figure 3 for $r = 1.15$ as an example) inside the unit square. For the distribution function of the random variable R we have

$$F(r) = \mathrm{P}(R \le r) = \begin{cases} \text{if } r < 0, \text{ then } 0 \\ \text{if } 0 \le r < 1, \text{ then } \dfrac{\pi}{4} r^2 \\ \text{if } 1 \le r < \sqrt{2}, \text{ then } \dfrac{\pi}{4} r^2 + \sqrt{r^2 - 1} - r^2 \arccos \dfrac{1}{r} \\ \text{if } \sqrt{2} \le r, \text{ then } 1 \end{cases} \tag{1}$$

Hence $\mathrm{P}(R \le 1) = \frac{\pi}{4}$ and the expected distance can be calculated from the integral

$\int_0^{\sqrt{2}} rf(r)\mathrm{d}r$, where $f = F'$ is the probability density function. Although this integral can be solved analytically and equals exactly $\frac{\sqrt{2}+\ln(1+\sqrt{2})}{3}$, we also can solve it numerically in an online integral calculator. Finally here should be mentioned that some parts of this exercise are beyond the high-school mathematics curriculum.

The next exercise is about the intriguing Monty-Hall probability puzzle about the concept of conditional probability, see e.g. https://en.wikipedia.org/wiki/Monty_Hall_problem. In this conceptual game a player is given a choice of three doors. Behind one door is a car, behind the other doors a goat. After the player picks one of the doors, the host of the game (who knows what lies behind each door) opens one of those doors having a goat and offers the player to switch his first choice or stay at the originally picked door. The player wins whatever is behind the door that he lets open after his final decision.

Exercise 4. Simulate many Monty-Hall games with an additional parameter $0 \leq p \leq 1$, which denotes the a priori probability that the player switches his door choice. Calculate the relative frequencies of each of the possible outcomes (stay,goat), (stay,car), (switch,goat) and (switch,car) and compare them to their corresponding probabilities! Calculate the conditional probability of winning a car if switching the first choice.

The tree graph on the left of Figure 4 shows a Monty-Hall game, where the rectangles are the doors (C=car, G=goat) and the border denotes the choice of the player. The probabilities of the outcomes are calculated by multiplying the probabilities on the corresponding branches. These theoretical formulae give rise to linear functions on the right diagram of Figure 4. In the

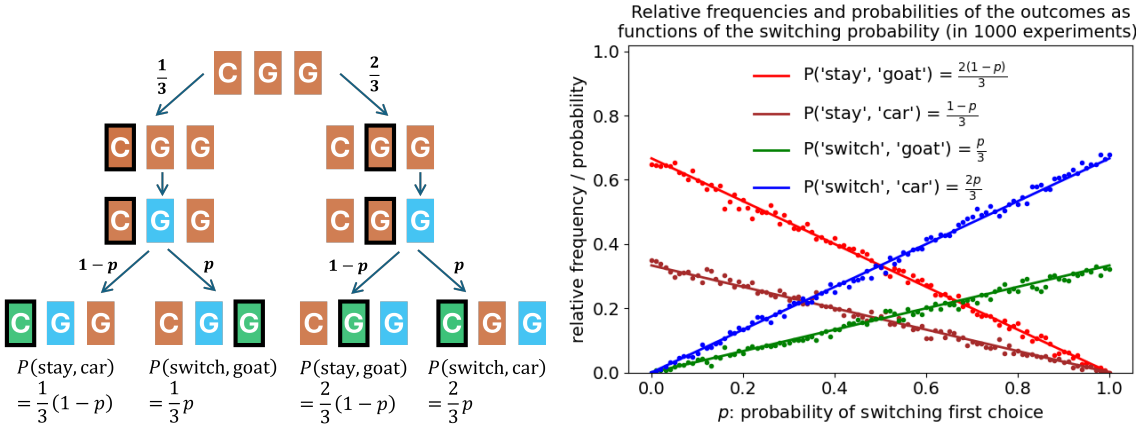

Figure 4. Monty-Hall puzzle

simulations we used the Python inbuilt pseudorandom number generator. Each point depicts the relative frequency result of 1000 games for the corresponding outcome. The experimental points fit fairly good onto the corresponding lines. Figure 5 shows the relative frequencies of the outcomes (on the left) and the conditional frequencies (on the right) tending to their corresponding probabilities as the number of random Monty-Hall games increases in case of the arbitrary chosen switching probability $p = 0.8$.

In the next exercise we utilize the PyOTP Python library for generating and verifying one-time passwords for implementing two-factor (2FA) authentication methods for applications which require users to log in. For more information and usage of the library see the Project description on https://pypi.org/project/pyotp/.

Exercise 5. Determine the probability distribution of the number of digits which occur at least twice in a randomly generated six digit long one time password (OTP for short) and calculate

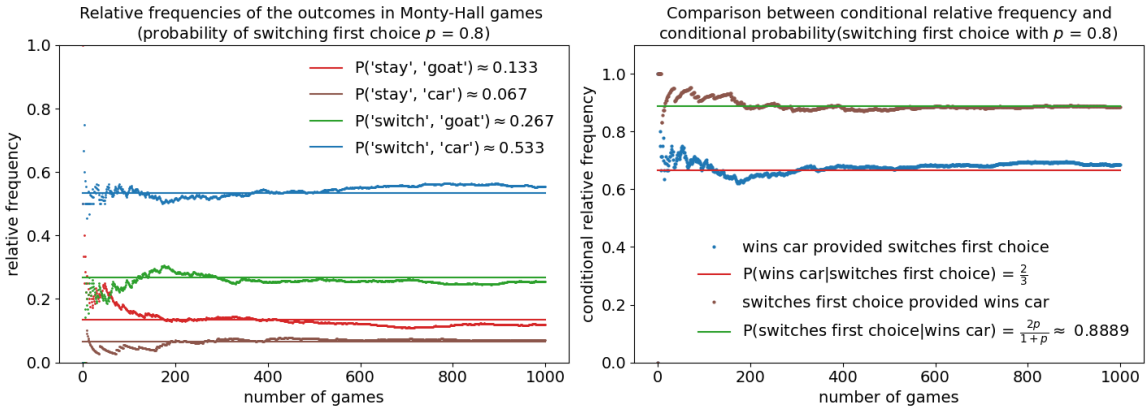

Figure 5. Monty-Hall simulation with $p = 0.8$

the expected value of this number! Generate many such OTPs and compare the experimental relative frequencies and averages with the theoretical probabilities and expected value!

Let the random variable X denote the number of digits which occur at least twice in the six digit long OTP. The possible values are $X = 0, 1, 2, 3$. Their probabilities can be calculated so that one chooses the appropriate number of distinct digit out of 10, then one chooses which digits how many times occur and finally one makes all possible permutations of the six prepared digits, c.f. Figure 6. The expected value of X from this distribution is $E(X) = 1.14265$. The experiments show that relative frequencies of the possible outcomes and the averages tend to the probabilities and to the expected value, respectively, c.f. the diagrams on Figure 6. The pseudorandomly generated OPT codes pretend true random behavior.

ABCDEF

$$P(X=0) = \frac{\binom{10}{6} \cdot 6!}{10^6} = 0{,}1512$$

AACDEF AAADEF AAAAEF AAAAAF AAAAAA

$$P(X=1) = \frac{\binom{10}{5} \cdot \binom{5}{1} \cdot \frac{6!}{2!} + \binom{10}{4} \cdot \binom{4}{1} \cdot \frac{6!}{3!} + \binom{10}{3} \cdot \binom{3}{1} \cdot \frac{6!}{4!} + \binom{10}{2} \cdot \binom{2}{1} \cdot \frac{6!}{5!} + \binom{10}{1} \cdot \binom{1}{1} \cdot \frac{6!}{6!}}{10^6} = 0{,}56575$$

AABBCD AAABBC AABBBB AAABBB

$$P(X=2) = \frac{\binom{10}{4} \cdot \binom{4}{2} \cdot \frac{6!}{2! \cdot 2!} + \binom{10}{3} \cdot \binom{3}{2} \cdot \binom{2}{1} \cdot \frac{6!}{2! \cdot 3!} + \binom{10}{2} \cdot \binom{2}{1} \cdot \frac{6!}{2! \cdot 4!} + \binom{10}{2} \cdot \frac{6!}{3! \cdot 3!}}{10^6} = 0{,}27225$$

AABBCC

$$P(X=3) = \frac{\binom{10}{3} \cdot \frac{6!}{2! \cdot 2! \cdot 2!}}{10^6} = 0{,}0108$$

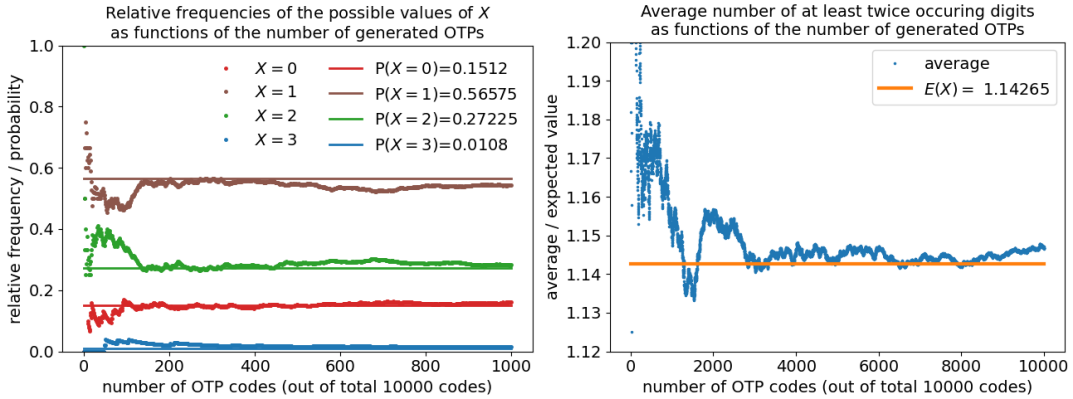

Figure 6. Distribution of at least twice occuring digits

In the last two exercises we examine the Hungarian lottery in which 5 winning numbers are drawn from 90. A comprehensive and continuously updated list of the drawings can be found e.g. at https://bet.szerencsejatek.hu/cmsfiles/otos.csv.

Exercise 6. Compute the probability of drawing a particular number in the hungarian 5 from 90 lottery! Let the random variable X denote how many times a particular number is drawn in a total of n mutually independent drawings! Determine the probability distribution of X and its expected value! Compare the theoretical results to the actual dataset of the drawn numbers!

The probability of the contrary event, that a particular number is not drawn equals $\frac{\binom{89}{5}}{\binom{90}{5}} = \frac{85}{90}$, hence the probability of drawing a particular number is $p = 1 - \frac{85}{90} = \frac{5}{90}$. Because the drawings are supposed to be mutually independent, the random variable X has a binomial distribution with success probability $p = \frac{5}{90}$ and n the total number of drawings, which is in our example $n = 3582$:

$$\mathrm{P}(X = k) = \binom{3582}{k} \left(\frac{5}{90}\right)^k \left(\frac{85}{90}\right)^{3582-k}, \text{ for } 0 \le k \le 3582.$$

The expected value of X is $E(X) = 3582 \cdot \frac{5}{90} = 199$. In order to compare this theoretical distribution to the actual drawing data, we count how many times each number was drawn. This gives a sample list of 90 numbers for the random variable X the histogram of which fits fairly good to the theoretical binomial distribution, c.f. Figure 7. Moreover, the expected value of X equals the average drawing frequency of the numbers.

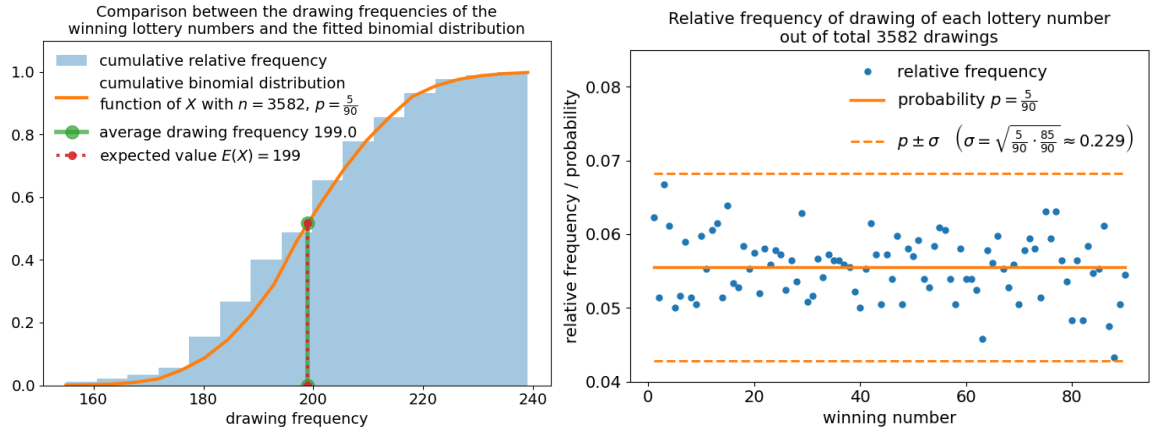

Figure 7. Drawing frequencies of the winning numbers

Exercise 7. In the 5 from 90 lottery the drawn winning numbers are published in ascending order. Let the random variable Y_ℓ ($\ell = 1; 2; 3; 4; 5$) denote the ℓ-th winning number. Determine the distribution of the random variables Y_ℓ! Calculate the expected value of each Y_ℓ! Compare the theoretical distributions to the actual dataset of the drawn numbers!

We utilize the hypergeometric distribution. The outcome $Y_\ell = k$ means that the first $\ell - 1$ winning numbers are drawn from the first $k - 1$ numbers, the ℓ-th winning number is k and the other $5 - \ell$ winning numbers are drawn form the other $90 - k$ numbers bigger than k. Hence there follows

$$\mathrm{P}(Y_\ell = k) = \frac{\binom{k-1}{\ell-1}\binom{90-k}{5-\ell}}{\binom{90}{5}}, \text{ for } \ell \le k \le 85 + \ell.$$

The expected value of Y_ℓ is determined by

$$E(Y_\ell) = \sum_{k=\ell}^{85+\ell} k \cdot \frac{\binom{k-1}{\ell-1}\binom{90-k}{5-\ell}}{\binom{90}{5}} = \ell \cdot \sum_{k=\ell}^{85+\ell} \frac{\binom{k}{\ell}\binom{90-k}{5-\ell}}{\binom{90}{5}} = \frac{91}{6}\ell.$$

These probability distributions and expected values can be compared to the relative frequency distributions and to the averages of the ℓ-th winning numbers: The values in Table 1 and the

ℓ	1	2	3	4	5
$E(Y_\ell) =$	15.16	30.3	45.5	60.6	75.83
average $\approx$	14.87	29.78	45.45	60.34	75.35

Table 1. Comparison of the average ℓ-th winning numbers with $E(Y_\ell)$

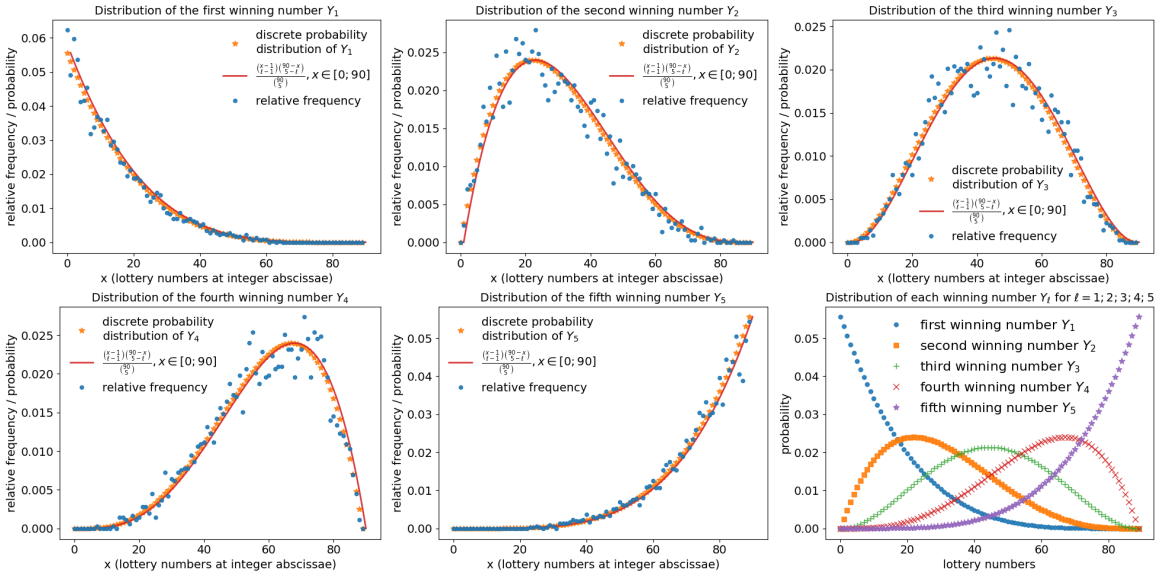

Figure 8. Distribution comparisons of the winning numbers

diagrams on Figure 8 show a very good accordance between the theoretical results and the actual dataset, which contains every drawing outcome from the first drawing in the 10th week of 1957 until the drawing in the 43rd week of 2025.

The distributions of the random variables Y_ℓ on Figure 8 can be extended to a function on the whole interval $x \in [0; 90]$ with $x \mapsto \frac{\binom{x-1}{\ell-1}\binom{90-x}{5-\ell}}{\binom{90}{5}}$, where the generalized binomial coefficients are defined with the Γ function as $\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)}$. For integer values of a and b this coincides with the usual definition because we have $n! = \Gamma(n+1)$ for every natural number n. The graphs of these functions and the diagrams of the distributions of Y_ℓ show a striking resemblance with the graphs of the Bernstein polynomials $B_{\ell,4}(x) = \binom{4}{\ell} x^\ell (1-x)^{4-\ell}$ for $\ell = 0; 1; 2; 3; 4$ resp., see e.g. [5]. We investigate this resemblance further in the next section.

3 Bernstein-like polynomials from the hypergeometric distribution

The Bernstein polynomials with parameters $n \geq 0$ and $0 \leq k \leq n$ are defined by

$$B_{k,n}(x) = \binom{n}{k} x^k (1-x)^{n-k}, \text{ for } 0 \leq x \leq 1. \tag{2}$$

These polynomials and their further generalized versions have an extensive literature and they found usage far beyond probability theory, c.f. [2]. Their connection to the binomial distribution $\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ with success probability $0 \leq p \leq 1$ is obvious by setting $p = x$. The binomial distribution describes the probability of a k-times success out of total n mutually independent trials by random sampling with replacement from a population which size is denoted by N. The success probability equals $p = \frac{K}{N}$, where $0 \leq K \leq N$ denotes the size of some subpopulation such that the choice of an element from this subpopulation is considered as a success.

If the random sampling is done without replacement, then we use the hypergeometric distribution instead, which is given by

$$\mathrm{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \text{ for } 0 \leq k \leq n, \tag{3}$$

where we also have $n \leq N$ and $k \leq K$. Defining here also $x = \frac{K}{N}$ as the fraction of the subpopulation size in the whole population, then using the properties of the generalized binomial coefficients we obtain alternative expressions for the hypergeometric distribution

$$\mathrm{P}(X = k) = \frac{\binom{xN}{k}\binom{(1-x)N}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k}\binom{N-n}{xN-k}}{\binom{N}{xN}}. \tag{4}$$

Herein we have the function $\frac{\binom{N-n}{xN-k}}{\binom{N}{xN}}$ of the variable x instead of $x^k(1-x)^{n-k}$, see (2). Moreover, if the parameters k, n, N are integers with $0 \leq k \leq n \leq N$, then the probabilities (4) as functions of the variable $0 \leq x \leq 1$ are in fact polynomials with these three parameters

$$B_{k,n,N}(x) = \binom{n}{k} \prod_{\ell=0}^{k-1} \frac{xN-\ell}{N-\ell} \prod_{\ell=k}^{n-1} \frac{(1-x)N-(\ell-k)}{N-\ell}, \text{ for } 0 \leq x \leq 1. \tag{5}$$

From the definition (5) we have $B_{0,0,N}(x) = B_{0,0}(x) = 1$, $B_{0,1,N}(x) = B_{0,1}(x) = 1-x$ and $B_{1,1,N}(x) = B_{1,1}(x) = x$ for each $N \geq 0$, i.e. the polynomials (2) and (5) coincide if $n = 0, 1$. Although this does not remain true for $n \geq 2$, we have the following

Theorem 1. *For each fixed parameter pair (k, n) and for $x \in [0; 1]$ we have*

$$\lim_{N\to\infty} B_{k,n,N}(x) = B_{k,n}(x). \tag{6}$$

Proof. For fixed parameters (k, n) with $0 \leq k \leq n \leq N$ the polynomials (2) consists of n factors for which we have

$$\lim_{N\to\infty} \frac{xN-\ell}{N-\ell} = x \text{ and } \lim_{N\to\infty} \frac{(1-x)N-(\ell-k)}{N-\ell} = 1-x$$

for each $0 \leq \ell \leq k-1$ and $k \leq \ell \leq n-1$, respectively. Hence the first k factors of (5) tend to x and the last $n-k$ factors of (5) tend to $1-x$, which implies (6). From the probability theory point of view this merely means that for sample sizes n way much smaller than the population size N it practically does not matter whether the random sampling occurs with or without replacement. □

Theorem 1 is visualized on Figure 9 for some arbitrary parameter values. Figure 9 also

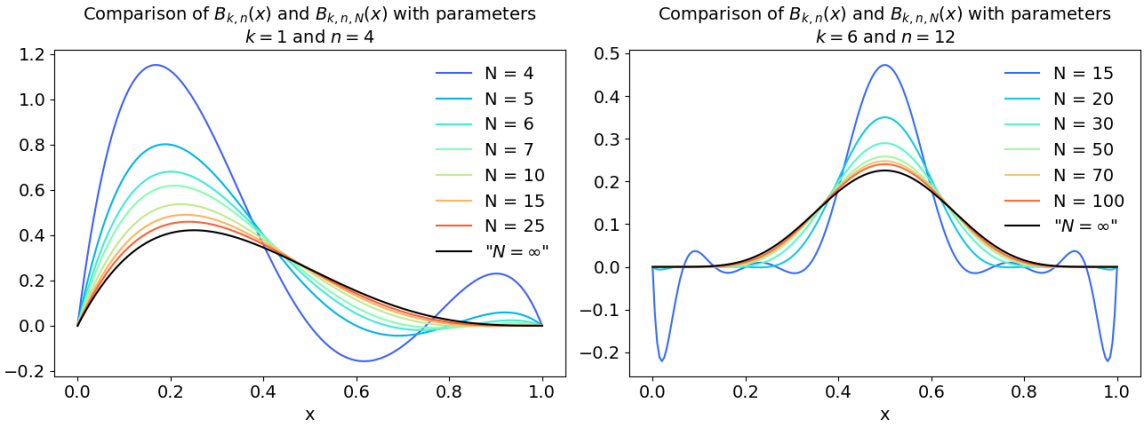

Figure 9. Illustration of Theorem 1

shows a major difference between (2) and (5), namely that whilst each of the polynomials (2) are nonnegative in the interval $[0;1]$ having multiple roots only on $x=0$ or $x=1$, the polynomials (5) have also negative values and only simple roots at the points $x=\frac{\ell}{N}$ for $0 \leq \ell \leq k-1$ or $x=1-\frac{\ell-k}{N}$ for $k \leq \ell \leq n-1$. Despite this difference the usual Berstein polynomials (2) derived from the binomial distribution and the ones (5) derived from the hypergeometric distribution have some other properties in common concerning for example symmetry or recursive relations between them. We formulate these in the following

Theorem 2. *The system* (5) *of polynomials has the following properties.*

- $B_{n-k,n,N}(1-x) = B_{k,n,N}(x)$
- $B_{k,n,N}(x) = \frac{n-k+1}{k} \cdot \frac{x-\frac{k-1}{N}}{1-x-\frac{n-k}{N}} B_{k-1,n,N}(x)$
- $B_{k,n,N}(x) = \frac{x-\frac{k-1}{N}}{1-\frac{n-1}{N}} \cdot B_{k-1,n-1,N}(x) + \frac{1-x-\frac{n-k-1}{N}}{1-\frac{n-1}{N}} \cdot B_{k,n-1,N}(x)$
- $\sum_{k=0}^{n} B_{k,n,N}(x) = 1$

Proof. Starting from the definition (4) the first identity follows as

$$B_{n-k,n,N}(1-x) = \frac{\binom{(1-x)N}{n-k}\binom{(1-(1-x))N}{n-(n-k)}}{\binom{N}{n}} = \frac{\binom{(1-x)N}{n-k}\binom{xN}{k}}{\binom{N}{n}} = B_{k,n,N}(x).$$

For the second and third identities we use the property $\binom{a}{b} = \frac{a-b+1}{b}\binom{a}{b-1}$ of the generalized binomial coefficients, which is proved by

$$\binom{a}{b} = \frac{\Gamma(a+1)}{\Gamma(b+1)\Gamma(a-b+1)} = \frac{\Gamma(a+1)}{b\Gamma(b)\frac{\Gamma(a-b+2)}{a-b+1}} = \frac{a-b+1}{b}\binom{a}{b-1},$$

where we used the functional identity $\Gamma(x+1) = x\Gamma(x)$ for the gamma function. From this follows also $\binom{a}{b} = \frac{b+1}{a-b}\binom{a}{b+1}$. Again starting with (4) we obtain

$$B_{k,n,N}(x) = \frac{\binom{(1-x)N}{n-k}\binom{xN}{k}}{\binom{N}{n}} = \frac{\frac{xN-k+1}{k}\binom{xN}{k-1}\cdot\frac{n-k+1}{(1-x)N-(n-k)}\binom{(1-x)N}{n-k+1}}{\binom{N}{n}}$$
$$= \frac{n-k+1}{k}\cdot\frac{x-\frac{k-1}{N}}{1-x-\frac{n-k}{N}}B_{k-1,n,N}(x),$$

which proves the second identity. The third identity is proved in two steps. On the one hand we have

$$\frac{x-\frac{k-1}{N}}{1-\frac{n-1}{N}}\cdot B_{k-1,n-1,N}(x) = \frac{xN-(k-1)}{N-(n-1)}\cdot\frac{\binom{xN}{k-1}\binom{(1-x)N}{n-k}}{\binom{N}{n-1}}$$
$$= \frac{xN-(k-1)}{N-(n-1)}\cdot\frac{\frac{k}{xN-(k-1)}\binom{xN}{k}\binom{(1-x)N}{n-k}}{\frac{n}{N-(n-1)}\binom{N}{n}} = \frac{k}{n}B_{k,n,N}(x).$$

On the other hand we also have

$$\frac{1-x-\frac{n-k-1}{N}}{1-\frac{n-1}{N}}\cdot B_{k,n-1,N}(x) = \frac{(1-x)N-(n-k-1)}{N-(n-1)}\cdot\frac{\binom{xN}{k}\binom{(1-x)N}{n-k-1}}{\binom{N}{n-1}}$$
$$= \frac{(1-x)N-(n-k-1)}{N-(n-1)}\cdot\frac{\binom{xN}{k}\frac{n-k}{(1-x)N-(n-k-1)}\binom{(1-x)N}{n-k}}{\frac{n}{N-(n-1)}\binom{N}{n}}$$
$$= \frac{n-k}{n}B_{k,n,N}(x).$$

Adding the latter two equalities the third identity follows

$$\frac{x-\frac{k-1}{N}}{1-\frac{n-1}{N}}\cdot B_{k-1,n-1,N}(x) + \frac{1-x-\frac{n-k-1}{N}}{1-\frac{n-1}{N}}\cdot B_{k,n-1,N}(x) = \frac{k}{n}B_{k,n,N}(x) + \frac{n-k}{n}B_{k,n,N}(x)$$
$$= B_{k,n,N}(x).$$

The last identity (partition of unity) expresses the mere fact that all the probabilities of the hypergeometric distribution denoted by $B_{k,n,N}(x)$ in (4) add up to 1. □

Remark 3. All the properties in Theorem 2 simplify to the respective properties of the ordinary Bernstein polynomials (2) with the same parameters k, n in the limiting case $N \to \infty$. The first (symmetric) and the third (recursive) property together give rise to a Pascal's triangle like structure of these polynomials (5) resembling the same structure of the Bernstein polynomials (2), however these have finite extent because the requirement $k \le n \le N$ limits the allowed values for the parameters k and n. This enriches a little bit the already vast literature of the generalizations of the Pascal triangle, see e.g. [3].

Conclusion

In this note we discussed a variety of examples how basic concepts of probability theory can be demonstrated for high-school students using big datasets. Obviously we only have scratched the surface of a huge multitude of possibilities since randomness can be found in datasets of several subjects taught in high-school. As a byproduct of an exercise concerning lottery number distributions we obtained Bernstein like polynomials from the hypergeometric distribution, some basic properties of which we briefly investigated.

Bibliography

[1] *Matematika részletes érettségi vizsgakövetelmény*, (Last updated 2021). URL: https://www.oktatas.hu/pub_bin/dload/kozoktatas/erettsegi/vizsgakovetelmenyek2024/matematika_2024_e.pdf.

[2] Lorentz, G.: *Bernstein Polynomials*, Toronto: University of Toronto Press, 1953.

[3] Németh, L. and Szalay, L.: *Áttekintés a hiperbolikus Pascal háromszögekről*, Dimenziók - Mathematical Notes, **8** (2020), No. 8, 61–74. doi: 10.20312/dim.2020.07.

[4] Wolfram-MathWorld: *Normal number*, (Last Updated: Oct 24 2025). URL: https://mathworld.wolfram.com/NormalNumber.html.

[5] Wolfram-MathWorld: *Bernstein polynomial*, (Last Updated: Oct 28 2025). URL: https://mathworld.wolfram.com/BernsteinPolynomial.html.

[6] Zsuppán, S.: *Érdekes matematikai problémák modellezése számítógéppel középiskolásoknak*, Dimenziók - Mathematical Notes, **11** (2023), No. 11, 77–82. doi: 10.20312/dim.2023.09.