Prace Naukowe Wyższej Szkoły Pedagogicznej w Częstochowie
Matematyka IX, Częstochowa 2003

Discrete Uniform Distribution and Paradoxes

Pavel Tlustý

Every teacher should make effort to motivate students for further studying. I have made good experience with using untypical examples and paradoxes.

In this article we can show some paradoxical situation associated with discrete uniform distribution. Uniform distribution is very simple and its properties are describe in all the textbook on probability. However, in some examples and situations we can see non-correct process or surprising statements.

Example 1:
In some probability books we can read the following problem.
Consider the set of all positive integers:

$$1,\ 2,\ 3,\ 4,\ 5,\ 6, \ldots$$

If you pick one of these number at random, what is probability that is multiple of 4 (it is silently supposed that all numbers have the same chance to be chosen)? Expected answer is $\frac{1}{4}$ because every fourth number is multiple of 4. Is it true?

Now rewrite the set of all positive integers in the form:

$$1,\ 2,\ 4,\ 3,\ 5,\ 8,\ 6,\ 7,\ 12,\ 9,\ 10,\ 16, \ldots$$

Every third number is multiple of 4 and therefore the answer is $\frac{1}{3}$ in this case. It is strange (probability depends on ordering).
The correct answer is simple. Since both sets are countably infinite, it's impossible to define discrete uniform distribution on these sets.

Example 2:
The empirical observation in many naturally occurring tables of numerical data (physical constants, populations, cost data, country area and so on) give an interesting conclusion: the leading significant (non-zero) digit is not uniformly distributed in $\{1, 2, \ldots, 8, 9\}$. But most people have the intuition that each of the digit $1, 2, \ldots, 8, 9$ are equally appear as the leading significant number. Why? Where is a mistake?

The first person who noticed this phenomenon was Simon Newcomb [2].

We tabulated the first digits of 199 physical constants [1] for demonstration of the phenomenon.

Leading digit	number of cases	percentage
1	67	33,6
2	39	19,5
3	21	10,7
4	16	8,1
5	17	8,5
6	14	7,1
7	7	3,5
8	8	4,0
9	10	5,0

Table 1: Empirical distribution of digits from physical constants table

How to explain this curios relationship?

Suppose that there really is a law of digit frequencies. Then the law should be universal, it does not depend on units. If we change the units, then proportions of digit frequencies should be the same - the law of digit frequencies should be scale invariant.

Let $\{D_1 = k\}$ be an event:

$$\{D_1 = k\} = \{\text{the first significant digit is } k\}.$$

Multiplication by 2, for example, converts all numbers starting with 5, 6, 7, 8 or 9 into numbers starting with 1. This implies that

$$P(D_1 = 1) = P(D_1 = 5) + P(D_1 = 6) + P(D_1 = 7) + \\ + P(D_1 = 8) + P(D_1 = 9) \qquad (*)$$

for scale-invariance under multiplication by 2 to hold, which is certainly not true that $P(D_1 = k)$ is the same for all $k \in \{1, 2, \ldots, 8, 9\}$.

Another explanation that the leading digit 1 should be more common than the other digits can be understood as follows: Start counting from 1: 1, 2, 3, ... As we reach 9, every digit will have been equally likely. From 10 to 19, we only have the leading digit 1. If we reach 99, all digits will be equally likely again and so on. 1 has always a lead, except for (9, 99, 999, ...).

We will find such a distribution that if we multiply all our numbers by arbitrary constants (as we do when change from metres to miles or euro to dollar and so on) then the distribution of first digit frequencies will be the same.

The area underneath density function between any two points a and b yields the probability of getting a value lying between a and b. Now we consider interval $\langle 1, 10)$. If in our case the density function is $f(x) = \frac{1}{x}$, then the relevant distribution is scale invariant. From integral calculus we know

$$\int_a^b \frac{1}{x}\,\mathrm{d}x = \ln b - \ln a.$$

For single digit numbers, if $a = n$ and $b = (n+1)$, then the probability that the digit equals n is $\ln \frac{n+1}{n}$. Because

$$\ln a \doteq 2,3026 \cdot \log_{10} a,$$

and our distribution should be scale invariant, we can use logs to the base 10 rather than natural logs with base e. Then

$$P(D_1 = 1) = \log_{10} 2 - \log_{10} 1 = \log_{10}\left(1 + \frac{1}{1}\right),$$

$$P(D_1 = 2) = \log_{10} 3 - \log_{10} 2 = \log_{10}\left(1 + \frac{1}{2}\right),$$

$$P(D_1 = 3) = \log_{10} 4 - \log_{10} 3 = \log_{10}\left(1 + \frac{1}{3}\right),$$

and so on. Then the significant-digit law is

$$P(D_1 = k) = \log_{10}\left(1 + \frac{1}{k}\right), \quad k = 1,\ 2,\ \ldots,\ 9.$$

It is easy to show that

$$\sum_{k=1}^{9} P(D_1 = k) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{k}\right) = 1.$$

This logarithmic distribution is the only distribution on the significant digits of real numbers which is invariant under changes of scale. This function leads to expected frequencies that are monotonically declining from about probability 0,3 for 1's down to about probability 0,05 for 9's (see table 2 and compare table 1).

Leading digit	probability
1	0,301
2	0,176
3	0,125
4	0,097
5	0,079
6	0,067
7	0,058
8	0,051
9	0,046

Table 2: Distribution of digits from significant-digit law

It is easy to show that this distribution is scale invariant. Multiplication by 2, for example, converts interval $\langle 1, 10)$ on interval $\langle 2, 20)$. Then

$$P(D_1 = 1) = \log_{10} 20 - \log_{10} 10 = \log_{10} \left(1 + \frac{1}{1}\right),$$

another probabilities $P(D_1 = 2),\ P(D_1 = 3), \ldots, P(D_1 = 9)$ stay the same.

We can generalize this law for the second and higher significant digits. Indicate

$$\{D_2 = k\} = \{\text{the second significant digit is } k\},$$

$$\{D_3 = k\} = \{\text{the third significant digit is } k\},$$

$$\{D_4 = k\} = \{\text{the fourth significant digit is } k\},$$

and so on.

The corresponding law for the second and higher significant digits is

$$P(D_1 = k_1, D_2 = k_2, \ldots, D_t = k_t) = \log_{10} \left[1 + \left(\sum_{i=1}^{t} k_i \times 10^{t-i}\right)^{-1}\right]$$

for $k_1 \in \{1, 2, \ldots, 9\}$ and $k_j \in \{0, 1, 2, \ldots, 9\}$, $j > 1$.

This says, for example, that the probability that the first three significant digits of a numbers are 1, 2, 5 is

$$P((D_1, D_2, D_3) = (1, 2, 5)) = \log_{10} \left(1 + \frac{1}{125}\right) \doteq 0,003\,461.$$

However

$$P((D_1, D_2, D_3) = (5, 2, 1)) = \log_{10} \left(1 + \frac{1}{521}\right) \doteq 0,000\,833.$$

References

[1] P. J. Mohr, B. N. Taylor, *CODATA Recommended value of the fundamental physical constants*, Journal of Physical and Chemical Reference Data, Vol. 28, No. 6, 1998.

[2] S. Newcomb, *Note on the frequency of use of the different digits in naturals numbers*, American Journal of Mathematics, 4, 39-40, 1881.

Department of Mathematics
Faculty of Education,
University České Budějovice,
Czech Republic