

Statistics - handout

Part-II: Discrete distributions

Andras Vetier

January 13, 2017

Contents

1	Discrete random variables and distributions	2
2	Uniform distribution (discrete)	5
3	Hyper-geometrical distribution	6
4	Binomial distribution	11
5	Geometrical distribution (pessimistic)	17
6	Geometrical distribution (optimistic)	19
7	Negative binomial distribution (pessimistic)	21
8	Negative binomial distribution (optimistic)	24
9	Poisson-distribution	26
10	Generating a random variable with a given discrete distribution	29
11	Mode of a distribution	30
12	Expected value of discrete distributions	33
13	Expected values of the most important discrete distributions	38
14	Expected value of a function of a discrete random variable	45
15	Moments of a discrete random variable	47
16	Average distance from the average, variance, standard deviation, etc.	49

1 Discrete random variables and distributions

When there are a finite or countably infinite number of outcomes, and each is assigned a (probability) value so that each value is non-negative and their sum is equal to 1, we say that a **discrete distribution** is given on the set of these outcomes. If x is an outcome, and the probability value assigned to it is denoted by $p(x)$, then the function p is called the **weight function** or **probability function** of the distribution. We emphasize that

$$p(x) \geq 0 \quad \text{for all } x$$

$$\sum_x p(x) = 1$$

The second property is called the **normalization property**.

The reader must have learned about the abstract notion of a **point-mass** in mechanics: certain amount of mass located in a certain point. This notion is really an abstract notion, because in reality, any positive amount of mass has a positive diameter, while the diameter of a point is zero. Although, point-masses in reality do not exist, our fantasy helps us to imagine them. It is advantageous to interpret the term $p(x)$ of a discrete distribution not only as the probability of the possible value x , but also as if an amount of mass $p(x)$ were located in the point x . Thus, a discrete distribution can be interpreted as a **point-mass distribution**.

When the possible results of an observation are real numbers, we say that we work with a **random variable**. Thus, to define a random variable, it means to refer to a (random) numerical value or ask a question so that the answer to the question means some (random) number. It is useful to abbreviate the declaration or question by a symbol (most often used are capital letters like X, Y, \dots , or Greek letters like α, β, \dots) so that later we can refer to the random variable by its symbol. Here are some examples for random variables:

1. Tossing a fair die, let X denote the random variable defined by

$$X = \text{the number which shows up on the top of the die}$$

or, equivalently,

$$X = \text{which number shows up on the top of the die?}$$

Then, the equality $X = 6$ means the event that *we toss a six with the die*, the inequality $X < 3$ means the event that *we toss a number which is less than 3*.

2. Tossing two coins, let Y denote the random variable defined by

$$Y = \text{the number of heads we get with the two coins}$$

When a random variable is thought of, we may figure out its **possible values**. The possible values of the random variable X defined above are 1, 2, 3, 4, 5, 6. The possible values of the random variable Y defined above are $\{0, 1, 2\}$. When there are a finite number or countably infinite number of outcomes, then we say that the distribution and the random variable are **discrete**. When we figure out the probability of each outcome of a discrete random variable, we say, that we figure out the **distribution of the random variable**. The distribution of the random variable X is quite simple: each of the numbers 1, 2, 3, 4, 5, 6 has the same probability, namely, $\frac{1}{6}$. This can be described, among others, by a table:

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

The distribution of the random variable Y is:

x	0	1	2
$p(x)$	0.25	0.50	0.25

Distributions are also described by formulas, as in the following chapters, where the most important discrete distributions will be listed.

Calculating a probability by summation. If an event corresponds to a subset A of the sample space, then the probability of the event can be calculated by the sum:

$$\mathbf{P}(A) = \sum_{x:x \in A} p(x)$$

In this sum, we summarize the probabilities of those outcomes which belong to the set A corresponding to the event.

Example 1: Will everybody play? Imagine that there is group of 10 people who like to play the card game called "bridge". Each evening, those of them who are free that evening come together in the house of one of them. As you probably know 4 persons are needed for this game. When all the 10 people come together, then 8 people can play, and 2 just stay there and watch the others playing. When only 9 people come together, then 8 people can play, and 1 just stays there and watches the others playing. When only 8 people come together, then all the 8 people can play. When only 7 people come together, then 4 people can play, and 3 just stay there and watch the others playing. And so on. Assume that the probability that exactly x people come together is $p(x)$, where $p(x)$ is given by the following table:

x	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.01	0.04	0.06	0.09	0.10	0.15	0.25	0.20	0.07	0.03

The question is: what is the probability that all the gathering people can play, that is, nobody has to only stay and watch the others playing? In other words: what is the probability that 4 or 8 gather together?

Solution. The solution is obvious: in order to get the answer, we have to add $p(4)$ and $p(8)$: $p(4) + p(8) = 0.09 + 0.20 = 0.29$.

Using Excel. The above summation was very easy. However, when there are many terms to add, then it may be convenient to perform the addition in Excel using simple summation, or using the `SUM`-command, or using the `SUMIF`-command, or using the `SUMPRODUCT`-command.

2 Uniform distribution (discrete)

A fair die has 6 possible values so that their probabilities are equal.

If the possible values of a random variable constitute an interval $[A, B]$ of integer numbers, and the probabilities of these values are equal, then the weight function is constant on this interval of integer numbers:

Weight function (probability function): It is a constant function. The value of the constant is the reciprocal of the number of integers in $[A, A + 1, \dots, B]$, which is $B - A + 1$.

$$p(x) = \frac{1}{B - A + 1} \quad \text{if } x = A, \dots, B$$

Here A and B are parameters:

A = left endpoint of the interval

B = right endpoint of the interval

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=A}^B \frac{1}{B - A + 1} = (B - A + 1) \frac{1}{B - A + 1} = 1$$

3 Hyper-geometrical distribution

Application:

Phenomenon: Red and green balls are in a box. A given number of draws are made *without replacement*.

Definition of the random variable:

X = the number of times we draw red

Parameters:

A = number of red balls in the box

B = number of green balls in the box

n = number of draws

Weight function (probability function):

$$p(x) = \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} \quad \text{if } \max(0, n-B) \leq x \leq \min(n, A)$$

Proof of the formula of the weight function. First let us realize that the possible values of X must satisfy the inequalities: $x \geq 0$, $x \leq n$, $x \leq A$, $n-x \leq B$. The last of these four inequalities means that $x \geq n-B$. Obviously, $x \geq 0$ and $x \geq n-B$ together mean that $x \geq \max(0, n-B)$, and $x \leq n$ and $x \leq A$ together mean that $x \leq \min(n, A)$. This is how we get that $\max(0, n-B) \leq x \leq \min(n, A)$. In order to find an expression for $p(x)$, we need to study the event " $X=x$ " which means that "there are exactly x red balls among the n balls" which we have drawn". In order to use the classical formula, first we realize that there are $\binom{A+B}{n}$ equally probable possible combinations. The favorable combinations are characterized by the property that x of the chosen balls are red, and $n-x$ are green. The number of such combinations is $\binom{A}{x} \binom{B}{n-x}$. The ratio of the two expressions gives the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=\max(0, n-B)}^{\min(n, A)} \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} = 1$$

We used the fact that

$$\sum_{x=\max(0,n-B)}^{\min(n,A)} \binom{A}{x} \binom{B}{n-x} = \binom{A+B}{n}$$

which can be derived by the following combinatorial argument: Assume that there are A red and B blue balls in a box. If n balls are drawn without replacement, then the number of combinations in which there are exactly x red balls is $\binom{A}{x} \binom{B}{n-x}$, so the total number of combinations is

$$\sum_{x=\max(0,n-B)}^{\min(n,A)} \binom{A}{x} \binom{B}{n-x}$$

On the other hand, the number of all possible combinations is obviously $\binom{A+B}{n}$.

Remark. Some textbooks define the same distribution by a different parameter setting, so that $N = A + B$, $K = A$ and n are considered parameters. Then the probability of x looks like as this:

$$\frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} \quad \text{if } \max(0, n - N + A) \leq x \leq \min(n, A)$$

In this approach, the parameters:

K = number of red balls in the box

N = number of all balls in the box

n = number of draws

Using Excel. In Excel, the function HYPGEOMDIST is associated to this second approach:

$$\frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} = \text{HYPGEOMDIST}(x; n; K; N)$$

In case of the first approach, the Excel-function HYPGEOMDIST should be used with the following parameter-setting:

$$\frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} = \text{HYPGEOMDIST}(x; n; A; A+B)$$

Here is an example for the application of the hyper-geometrical distribution:

Example. Lottery. There are two popular lotteries in Hungary. One is called 5-lottery, the other is called 6-lottery. On a 5-lottery ticket, the player has to fill in 5 numbers out of 90, and on Sunday evening 5 numbers are chosen at random out of the 90 numbers. If the chosen numbers are the same as the filled numbers, then the player has 5 hits, and wins a huge amount of money. If 4 of the chosen numbers are among the filled numbers, then the player has 4 hits, and wins a good amount of money. 3 hits mean a not negligible amount of money. 2 hits mean a some small amount of money. For 1 hit or 0 hits, the player does not get any money. If you play this lottery, you will be interested in knowing how much is the probability of each of the possible hits. On a 6-lottery ticket, the player has to fill in 6 numbers out 45 of , and on Saturday evening 6 numbers are chosen at random out of the 45 numbers. We give the probability of each of the possible hits for the 6-lottery, as well.

Solution. The answers are given by the formula of the hyper-geometrical distribution.
5-lottery:

$$P(5 \text{ hits}) = \frac{\binom{5}{5} \binom{85}{0}}{\binom{90}{5}} = 0,00000002 = 2 \cdot 10^{-8}$$

$$P(4 \text{ hits}) = \frac{\binom{5}{4} \binom{85}{1}}{\binom{90}{5}} = 0,00000967 = 1 \cdot 10^{-5}$$

$$P(3 \text{ hits}) = \frac{\binom{5}{3} \binom{85}{2}}{\binom{90}{5}} = 0,00081230 = 2 \cdot 10^{-4}$$

$$P(2 \text{ hits}) = \frac{\binom{5}{2} \binom{85}{3}}{\binom{90}{5}} = 0,02247364 = 2 \cdot 10^{-2}$$

$$P(1 \text{ hit}) = \frac{\binom{5}{1} \binom{85}{4}}{\binom{90}{5}} = 0,23035480 = 2 \cdot 10^{-1}$$

$$\mathbf{P}(0 \text{ hits}) = \frac{\binom{5}{0} \binom{85}{5}}{\binom{90}{5}} = 0,74634956 = 7 \cdot 10^{-1}$$

6-lottery:

$$\mathbf{P}(6 \text{ hits}) = \frac{\binom{6}{6} \binom{39}{0}}{\binom{45}{6}} = 0,00000012 = 1 \cdot 10^{-7}$$

$$\mathbf{P}(5 \text{ hits}) = \frac{\binom{6}{5} \binom{39}{1}}{\binom{45}{6}} = 0,00002873 = 3 \cdot 10^{-5}$$

$$\mathbf{P}(4 \text{ hits}) = \frac{\binom{6}{4} \binom{39}{2}}{\binom{45}{6}} = 0,00136463 = 1 \cdot 10^{-3}$$

$$\mathbf{P}(3 \text{ hits}) = \frac{\binom{6}{3} \binom{39}{3}}{\binom{45}{6}} = 0,02244060 = 2 \cdot 10^{-2}$$

$$\mathbf{P}(2 \text{ hits}) = \frac{\binom{6}{2} \binom{39}{4}}{\binom{45}{6}} = 0,15147402 = 2 \cdot 10^{-1}$$

$$\mathbf{P}(1 \text{ hit}) = \frac{\binom{6}{1} \binom{39}{5}}{\binom{45}{6}} = 0,42412726 = 4 \cdot 10^{-1}$$

$$\mathbf{P}(0 \text{ hits}) = \frac{\binom{6}{0} \binom{39}{6}}{\binom{45}{6}} = 0,40056464 = 4 \cdot 10^{-1}$$

4 Binomial distribution

Applications:

1. Phenomenon: Red and green balls are in a box. A given number of draws are made *with replacement*.

Definition of the random variable:

X = the number of times we draw red

Parameters:

n = number of draws

p = probability of drawing red at one draw = $\frac{\text{number of red}}{\text{number of all}}$

2. Phenomenon: We make a given number of experiments for an event.

Definition of the random variable:

X = number of times the event occurs

Parameters:

n = number of experiments

p = probability of the event

3. Phenomenon: A given number of independent events which have the same probability are observed.

Definition of the random variable:

X = how many of the events occur

Parameters:

n = number of events

p = common probability value of the events

Weight function (probability function):

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{if } x = 0, 1, 2, \dots, n$$

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event " $X=x$ " which means that "the number times we draw red is

x ", which automatically includes that the number of times we draw green is $n - x$. If the variation of the colors were prescribed, for example, we would prescribe that the 1st, the 2nd, and so on the x th should be red, and the $(x + 1)$ th, the $(x + 2)$ th, and so on the n th should be green, then the probability of each of these variations would be $p^n(1 - p)^{n-x}$. Since there are $\frac{n!}{x!(n-x)!} = \binom{n}{x}$ variations, the the product of $\binom{n}{x}$ and $p^n(1 - p)^{n-x}$ really yields the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = (p + (1-p))^n = 1^n = 1$$

We used the binomial formula

$$\sum_{x=0}^n \binom{n}{x} a^x b^{n-x} = (a + b)^n$$

known from algebra, with $a = p$, $b = 1 - p$.

Approximation with hyper-geometrical distribution: Let n and p be given numbers. If A and B are large, and $\frac{A}{A+B}$ is close to p , then the terms of the hyper-geometrical distribution with parameters A, B, n approximate the terms of the binomial distribution with parameters n and p :

$$\frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} \approx \binom{n}{x} p^x (1-p)^{n-x}$$

More precisely: for any fixed n, p and x , $x = 0, 1, 2, \dots$, it is true that if $A \rightarrow \infty$, $B \rightarrow \infty$, $\frac{A}{A+B} \rightarrow p$, then

$$\frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} \rightarrow \binom{n}{x} p^x (1-p)^{n-x}$$

Proof of the approximation is left for the interested reader as a limit-calculation exercise.

Remark. If we think of the real-life application of the hyper-geometrical and binomial distributions, then the statement becomes quite natural: if we draw a given (small) number of times from a box which contains a large number of red and blue balls, then the fact whether we draw without replacement (which would imply hyper-geometrical

distribution) or with replacement (which would imply binomial distribution) has only a negligible effect.

Using Excel. In Excel, the function BINOMDIST is associated to this distribution. If the last parameter is FALSE, we get the weight function of the binomial distribution:

$$\binom{n}{x} p^x (1-p)^{n-x} = \text{BINOMDIST}(x; n; p; \text{FALSE})$$

If the last parameter is TRUE, then we get the so called accumulated probabilities for the binomial distribution:

$$\sum_{x=0}^k \binom{n}{x} p^x (1-p)^{n-x} = \text{BINOMDIST}(k; n; p; \text{TRUE})$$

Example 1. Air-plane tickets. Assume that there are 200 seats on an air-plane, and 202 tickets are sold for a flight on that air-plane. If some passengers - for different causes - miss the flight, then there remain empty seats on the air-plane. This is why some air-lines sell more tickets than the number of seats in the air-plain. Clearly, if 202 tickets are sold, then it may happen that more people arrive at the gate of the flight at the air-port than 200, which is a bad situation for the air-line. Let us assume that each passenger may miss the flight independently of the others with a probability $p = 0.03$. If $n = 202$ tickets are sold, then how much is the probability that there are more than 200 people?

Solution. The number of occurring follows, obviously, binomial distribution with parameters $n = 202$ and $p = 0.03$.

$$\mathbf{P}(\text{More people occur than 200}) =$$

$$\mathbf{P}(0 \text{ or } 1 \text{ persons miss the flight}) =$$

$$\text{BINOMDIST}(1; 202; 0,03; \text{TRUE}) \approx 0,015$$

We see that under the assumptions, the bad situation for the air-line will take place only in 1-2 % of the cases.

Using Excel. It is important for an air-line which uses this strategy to know how the bad situation depends on the parameters n and p . The answer is easily given by an Excel formula:

$$\text{BINOMDIST}(n-201; n; p; \text{TRUE})$$

Using this formula, it is easy to construct a table in Excel which expresses the numerical values of the probability of the bad situation in terms of n and p :

Example 2. How many chairs? Let us assume that each of the 400 students at a university attends a lecture independently of the others with a probability 0.6. First, let us assume that there are, say, only 230 chairs in the lecture-room. If more than 230 students attend, then some of the attending students will not have a chair. If 230 or less students attend, then all attending students will have a chair. The probability that the second case holds:

$$P(\text{All attending students will have a chair}) =$$

$$P(230 \text{ or less students attend}) =$$

$$\text{BINOMDIST}(230; 400; 0,6; \text{TRUE}) \approx 0,17$$

Now, let us assume that there are 250 chairs. If more than 250 students attend, then some of the students will not have a chair. Now:

$$P(\text{All attending students will have a chair}) =$$

$$P(250 \text{ or less students attend}) =$$

$$\text{BINOMDIST}(250; 400; 0,6; \text{TRUE}) \approx 0,86$$

We may want to know: how many chairs are needed to guarantee that

$$P(\text{All attending students will have a chair}) \geq 0,99$$

Remark. The following wrong argument is quite popular among people who have not learnt probability theory. Clearly, if there are 400 chairs, then :

$$P(\text{all attending students will have a chair}) = 1$$

So, they think, taking the 99 % of 400, the answer is 396. We will see that much less chairs are enough, so 396 chairs would be a big waste here.

Solution. To give the answer we have to find c so that

$$P(\text{All attending students will have a chair}) =$$

$$P(c \text{ or less students attend}) =$$

$$\text{BINOMDIST}(c; 400; 0,6; \text{TRUE}) \geq 0,99$$

Using Excel, we may easily construct a table for $\text{BINOMDIST}(c; 400; 0,6; \text{TRUE})$.

A part of the table is printed here:

c	P(all attending students will have a chair)
260	0,9824
261	0,9864
262	0,9897
263	0,9922
264	0,9942
265	0,9957

We see that if $c < 263$, then

$$P(\text{All attending students will have a chair}) < 0,99$$

if $c \geq 263$, then

$$P(\text{All attending students will have a chair}) \geq 0,99$$

Thus, we may conclude that 263 chairs are enough.

Remark. The way how we found the value of c was the following: we went down in the second column on the table, and when we first found a number greater than or equal to 0.99, we took the c value standing there in the first column.

Using Excel. In Excel, there is a special command to find the value c in such problems: $\text{CRITBINOM}(n; p; y)$ gives the smallest c value for which $\text{BINOMDIST}(c; n; p; \text{TRUE}) \geq y$. Specifically, as you may be convinced

$$\text{CRITBINOM}(400; 0,6; 0,99) = 263$$

Using the $\text{CRITBINOM}(n; p; y)$ -command, we get that

y	CRITBINOM(400 ; 0,6 ; y)
0,9	253
0,99	263
0,999	270
0,9999	276
0,99999	281
0,999999	285

which shows, among others, that with 285 chairs:

$$P(\text{all attending students will have a chair}) \geq 0,999999$$

Putting only 285 chairs instead of 400 into the lecture-room, we may save 125 chairs on the price of a risk which has a probability less than 0,0000001. Such facts are important when the size or capacity of an object is planned.

Example 3. Computers and viruses. There are 12 computers in an office. Each of them, independent of the others, has a virus with a probability 0.6. Each computer which has a virus still works with a probability 0.7, independent of the others. The number of computers having a virus is a random variable V . It is obvious that V has a binomial distribution with parameters 12 and 0.6. The number of computers having a virus, but still working is another random variable, which we denote by W . It is obvious that if $V = i$, then W has a binomial distribution with parameters i and 0.7. It is not difficult to see that W has a binomial distribution with parameters 12 and $(0.6)(0.7) = 0.42$. In the following file, we simulate V and W , and first calculate the following probabilities:

$$P(V = 4)$$

$$P(W = 3|V = 4)$$

$$P(V = 4 \text{ and } W = 3)$$

Then we calculate the more general probabilities:

$$P(V = i)$$

$$P(W = j|V = i)$$

$$P(V = i \text{ and } W = j)$$

Finally, we calculate the probability

$$P(W = j)$$

in two ways: first from the probabilities

$$P(V = i \text{ and } W = j)$$

by summation, and then using the BINOMDIST Excel function with parameters 12 and 0.42. You can see that we get the same numerical values for

$$P(W = j)$$

in both ways.

Example 4. Analyzing the behavior of the relative frequency. If we make 10 experiments for an event which has a probability 0.6, then the possible values of the frequency of the event (the number of times it occurs) are the numbers 0, 1, 2, . . . , 10, and the associated probabilities come from the binomial distribution with parameters 10 and 0.6. The possible values of the relative frequency of the event (the number of times it occurs divided by 10) are the numbers 0.0, 0.1, 0.2, . . . , 1.0, and the associated probabilities are the same: they come from the binomial distribution with parameters 10 and 0.6. We may call this distribution as a **compressed binomial distribution**.

The special case of of the binomial distribution when $n = 1$ has a special name:

Indicator distribution with parameter p

Application:

Phenomenon: An event is considered. We perform an experiment for the event.

Definition of the random variable:

$$X = \begin{cases} 0 & \text{if the event does not occur} \\ 1 & \text{if the event occurs} \end{cases}$$

Parameter:

$$p = \text{probability of the event}$$

Weight function (probability function):

$$p(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

5 Geometrical distribution (pessimistic)

Remark. The adjective *pessimistic* will become clear when we introduce and explain the meaning of an *optimistic* geometrical distribution as well.

Applications:

1. Phenomenon: There are red and green balls in a box. We make draws *with replacement* until we draw the first red ball.

Definition of the random variable:

X = how many green balls are drawn before the first red

Parameter:

p = the probability of red at each draw

2. Phenomenon: We make experiments for an event until the first occurrence of the event (until the first "success").

Definition of the random variable:

X = the number of experiments needed before the first occurrence

or, with the other terminology,

X = the number failures before the first success

Parameter:

p = probability of the event

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

X = the number of non-occurrences before the first occurrence

Parameter:

p = common probability value of the events

Weight function (probability function):

$$p(x) = (1 - p)^x p \quad \text{if } x = 0, 1, 2, \dots$$

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event " $X=x$ " which means that "the before the first red ball there are x green balls", which means that the 1st draw is green, and the 2nd draw is green, and the 3rd draw is green, and so on the x th draw is green, and the $(x + 1)$ th draw is red. The probability of this is equal to $(1 - p)^x p$, which is the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} (1 - p)^x p = \frac{p}{1 - (1 - p)} = \frac{p}{p} = 1$$

We used the summation formula for infinite geometrical series verbalized as "First term divided by one minus the quotient":

$$\sum_{x=0}^n q^x a = \frac{a}{1 - q} = \frac{p}{p} = 1$$

Using Excel. In Excel, there is no a special function associated to this distribution. However, using the power function POWER and multiplication, it is easy to construct a formula for this distribution:

$$(1 - p)^x p = \text{POWER}(1-p; x) * p$$

6 Geometrical distribution (optimistic)

Applications:

1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the first red ball

Definition of the random variable:

X = how many draws are needed until the first red

Parameter:

p = the probability of red at each draw

2. Phenomenon: We make experiments for an event until the first occurrence of the event.

Definition of the random variable:

X = the number of experiments needed until the first occurrence

Parameter:

p = probability of the event

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

X = the rank of the first occurring event in the sequence

Parameter:

p = common probability value of the events

Weight function (probability function):

$$p(x) = (1 - p)^{x-1}p \quad \text{if } x = 1, 2, \dots$$

Proof of the formula of the weight function. In order to find an expression for $p(x)$, we need to study the event " $X=x$ " which means that "the first red ball occurs at the x th draw", which means that the 1st draw is green, and the 2nd draw is green, and the 3rd draw is green, and so on the $(x - 1)$ th draw is green, and the x th draw is red. The probability of this is equal to $(1 - p)^{x-1}p$, which is the formula of the weight function.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=1}^{\infty} (1-p)^{x-1} p = \frac{p}{1-(1-p)} = \frac{p}{p} = 1$$

We used the summation formula for infinite geometrical series verbalized as "First term divided by one minus the quotient":

$$\sum_{x=0}^n q^x a = \frac{a}{1-q} = \frac{p}{p} = 1$$

Using Excel. In Excel, there is no a special function associated to this distribution. However, using the power function POWER and multiplication, it is easy to construct a formula for this distribution:

$$(1-p)^{x-1} p = \text{POWER}(1-p; x-1) * p$$

Remark. The terms **pessimistic** and **optimistic** are justified by the attitude that drawing a red ball at any draw may be interpreted as a success, drawing a green ball at any draw may be interpreted as a failure. Now, a person interested in the number of draws until *the first success* can be regarded as an optimistic person compared to someone else who is interested in the *number of failures* before the first success.

7 Negative binomial distribution (pessimistic)

Applications:

1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the r th red.

Definition of the random variable:

X = how many green balls are drawn before the r th red ball

Parameters:

r = the number of times we want to pick red

p = the probability of drawing red at each draw

2. Phenomenon: We make experiments for an event until the r th occurrence of the event (until the r th "success").

Definition of the random variable:

X = the number of non-occurrences before the r th occurrence

or, with the other terminology,

X = the number of failures before the r th success

Parameters:

r = the number of times we want to pick red

p = the probability of the event

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

X = the number of non-occurrences before the r th occurrence

or, with the other terminology,

X = the number of failures before the r th success

Parameters:

r = the number of times we want occurrence

p = common probability value of the events

Weight function (probability function):

$$p(x) = \binom{x+r-1}{x} p^r (1-p)^x \quad \text{if } x = 0, 1, 2, \dots \quad (\text{combinatorial form})$$

$$p(x) = \binom{-r}{x} p^r (-(1-p))^x \quad \text{if } x = 0, 1, 2, \dots \quad (\text{analytical form})$$

Proof of the combinatorial form. In order to find an expression for $p(x)$, we need to study the event " $X=x$ " which means that "before the r th red ball, we draw exactly x green balls". This means that among the first $x+r-1$ draws there are exactly x green balls, and the $(x+r)$ th draw is a red. The probability that among the first $x+r-1$ draws there are exactly x green balls is equal to

$$\binom{x+r-1}{x} (1-p)^x p^{r-1}$$

and probability that the $(x+r)$ th draw is a red is equal to p . The product of these two expressions yields the combinatorial form of the weight function.

Proof of the analytical form. We derive it from the combinatorial form. We expand the binomial coefficient into a fraction of products, and we get that

$$\begin{aligned} \binom{x+r-1}{x} &= \frac{(x+r-1)(x+r-2)\dots(r+1)(r)}{x!} = \\ &= \frac{(-(x+r-1))(-(x+r-2))\dots(-(r+1))(-r)}{x!} (-1)^x = \\ &= \frac{(-r)(-(r+1))\dots(-(x+r-3))(-(x+r-1))}{x!} (-1)^x = \\ &= \frac{(-r)(-r-1)\dots(-r-(x-2))(-r-(x-1))}{x!} (-1)^x = \\ &= \binom{-r}{x} (-1)^x \end{aligned}$$

If both the leftmost and the rightmost side of this equality are multiplied by $p^r(1-p)^x$, we get the combinatorial form on the left side and the analytical form on the right side.

Remark. Since the analytical form of the weight function contains the negative number $-r$ in the upper position of the binomial coefficient, the name "negative binomial with parameter r " is used for this distribution.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} p(x) = 1$$

We used the summation formula

Using Excel. In Excel, the function `NEGBINOMDIST` gives the individual terms of this distribution:

$$\binom{x+r-1}{r-1} p^r (1-p)^x = \text{NEGBINOMDIST}(x; r; p)$$

This Excel function does not offer a `TRUE`-option to calculate the summarized probabilities. The summarized probability

$$\sum_{i=0}^k \binom{i+r-1}{r-1} p^r (1-p)^i$$

can be calculated, obviously, by summation. However, using a trivial relation between negative binomial and binomial distribution, the summarized probability can be directly calculated by the Excel formula `1-BINOMDIST(r-1; k; p; TRUE)`

8 Negative binomial distribution (optimistic)

Applications:

1. Phenomenon: Red and green balls are in a box. We make draws with replacement until we draw the r th red.

Definition of the random variable:

X = how many draws are needed until the r th red

Parameters:

r = the number of times we want to pick red

p = the probability of drawing red at each draw

2. Phenomenon: We make experiments for an event until the r th occurrence of the event.

Definition of the random variable:

X = the number of experiments needed until the r th occurrence

Parameters:

r = the number of times we want occurrence

p = probability of the event

3. Phenomenon: An infinite sequence of independent events which have the same probability is considered.

Definition of the random variable:

X = the rank of the r th occurring event in the sequence

Parameters:

r = the number of times we want occurrence

p = common probability value of the events

Weight function (probability function):

$$p(x) = \binom{x-1}{x-r} p^r (1-p)^{x-r} \quad \text{if } x = r, r+1, r+2, \dots$$

Using Excel. In Excel, the function `NEGBINOMDIST` can be used for this distribution. However, keep in mind that the function `NEGBINOMDIST` is directly associated to the **pessimistic** negative binomial distribution, so for the **optimistic** negative binomial distribution we have to use the function `NEGBINOMDIST` with the following parameter-setting:

$$\binom{x-1}{r-1} p^r (1-p)^{x-r} = \text{NEGBINOMDIST}(x-r; r; p)$$

9 Poisson-distribution

Applications:

1. Phenomenon: We make a large number of experiments for an event which has a small probability.

Definition of the random variable:

$$X = \text{number of times the event occurs}$$

Parameter:

$$\lambda = \text{the theoretical average of the number of the times the event occurs}$$

Remark. If the probability of the event is p , and we make n experiments, then

$$\lambda = np$$

2. Phenomenon: Many, independent events which have small probabilities are observed.

Definition of the random variable:

$$X = \text{how many of them occur}$$

Parameter:

$$\lambda = \text{the theoretical average of the number of the occurring events}$$

Remark. If the number of events is n , and each event has the same probability p , then

$$\lambda = np$$

Weight function (probability function):

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{if } x = 0, 1, 2, \dots$$

Approximation with binomial distribution: if n is large and p is small, then the terms of the binomial distribution with parameters n and p approximate the terms of the Poisson distribution with parameter $\lambda = np$:

$$\binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\lambda^x}{x!} e^{-\lambda}$$

More precisely: for any fixed λ and x so that $\lambda > 0$, $x = 0, 1, 2, \dots$, it is true that if $n \rightarrow \infty$, $p \rightarrow 0$ so that $np \rightarrow \lambda$, then

$$\binom{n}{x} p^x (1-p)^{n-x} \rightarrow \frac{\lambda^x}{x!} e^{-\lambda}$$

Proof of the approximation.

$$\binom{n}{x} p^x (1-p)^{n-x} = \frac{n(n-1)(n-2)\dots(n-(x-1))}{x!} p^x (1-p)^{n-x} =$$

$$\frac{\left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \left(\frac{n-2}{n}\right) \dots \left(\frac{n-(x-1)}{n}\right)}{x!} (np)^x \frac{(1-p)^n}{(1-p)^n} \rightarrow$$

$$\frac{(1)(1)(1)\dots(1)}{x!} (\lambda)^x \frac{e^{-\lambda}}{1} = \frac{\lambda^x}{x!} e^{-\lambda}$$

We used the fact that

$$(1-p)^n \rightarrow e^{-\lambda}$$

which follows from the well-known calculus rule stating that if $u \rightarrow 1$ and $v \rightarrow \infty$, then $\lim u^v = e^{\lim uv}$.

Proof of the normalization property.

$$\sum_x p(x) = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$$

We used the fact known from the theory of Taylor-series that

$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$$

Using Excel. In Excel, the function POISSON is associated to this distribution. If the last parameter is FALSE, we get the weight function of the Poisson-distribution:

$$\frac{\lambda^x}{x!} e^{-\lambda} = \text{POISSON}(x; \lambda; \text{FALSE})$$

If the last parameter is TRUE, then we get the so called accumulated probabilities for the Poisson-distribution:

$$\sum_{x=0}^k \frac{\lambda^x}{x!} e^{-\lambda} = \text{POISSON}(k; \lambda; \text{TRUE})$$

Example. How many fish? Some years ago I met an old fisherman. He was fishing in a big lake, in which many small fish were swimming regardless of each other. He raised his net from time to time, and collected the fish if there were any. He told me that out of 100 cases the net is empty only 6 times or so, and then he added: "If you can guess the number of fish in the net when I raise it out of the water the next time, I will give you a big sack of money." I am sure he would not have said such a promise if he knew that his visitor was a well educated person in probability theory! I was thinking a little bit, then I made some calculation, and then I said a number. Which number did I say?

Solution. I started to think like this: The number of fish in the net is a random variable. Since the number of fish in the lake is large, and for each fish the probability of being caught is approximately equal to the area of the net compared to the area of the lake, which is a small value, and the fish swim independently, this random variable follows a Poisson distribution. The information that "out of 100 cases the net is empty only 6 times or so" means that the probability of 0 is 6/100. The formula for the Poisson distribution at 0 is $\frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$, so $e^{-\lambda} = 6/100$, from which we get $\lambda = \ln(100/6) \approx 2.8$. This means that the number of fish in the net follows the Poisson distribution with parameter 2.8.

The numerical values of this distribution are calculated:

x	0	1	2	3	4	5	6	7	8	9
p()	0,06	0,17	0,24	0,22	0,16	0,09	0,04	0,02	0,01,	0,00

We see that the most probable value is the number 2. So, I said "2".

10 Generating a random variable with a given discrete distribution

It is an important and useful fact that a random variable with a given discrete distribution easily can be generated by a calculator or a computer. The following is a method for this.

Assume that a discrete distribution is given:

x_1	x_2	x_3	x_4	...
p_1	p_2	p_3	p_4	...

We may calculate the so called **accumulated probabilities**:

$$P_0 = 0$$

$$P_1 = P_0 + p_1 = p_1$$

$$P_2 = P_1 + p_2 = p_1 + p_2$$

$$P_3 = P_2 + p_3 = p_1 + p_2 + p_3$$

$$P_4 = P_3 + p_4 = p_1 + p_2 + p_3 + p_4$$

...

These probabilities clearly constitute an increasing sequence between 0 and 1. So the following definition of the random variable X based on a random number RND is correct, and obviously guarantees that the distribution of the random variable X is the given discrete distribution:

$$X = \begin{cases} x_1 & \text{if } P_0 < \text{RND} < P_1 \\ x_2 & \text{if } P_1 < \text{RND} < P_2 \\ x_3 & \text{if } P_2 < \text{RND} < P_3 \\ x_4 & \text{if } P_3 < \text{RND} < P_4 \\ \dots & \dots \end{cases}$$

11 Mode of a distribution

The most probable value (values) of a discrete distribution is (are) called the **mode** (**modes**) of the distribution. The following method is applicable in calculating the mode of a distribution in many cases.

Method to determine the mode. Assume that the possible values of a distribution constitute an interval of integer numbers. For an integer x , let the probability of x be denoted by $p(x)$. The mode is that value of x for which $p(x)$ is maximal. Finding the maximum by the method, known from calculus, of taking the derivative, equate it to 0, and then solving the arising equation is not applicable, since the function $p(x)$ is defined only for integer values of x , and thus, differentiation is meaningless for $p(x)$. However, let us compare two adjacent function values, for example $p(x - 1)$ and $p(x)$ to see which of them is larger than the other:

$$p(x - 1) < p(x) \quad \text{or} \quad p(x - 1) = p(x) \quad \text{or} \quad p(x - 1) > p(x)$$

In many cases, after simplifications, it turns out that there exists a real number c so that the above inequalities are equivalent to the inequalities

$$x < c \quad \text{or} \quad x = c \quad \text{or} \quad x > c$$

In other words:

$$\begin{aligned} \text{when } x < c, & \quad \text{then } p(x - 1) < p(x), \\ \text{when } x = c, & \quad \text{then } p(x - 1) = p(x), \\ \text{when } x > c, & \quad \text{then } p(x - 1) > p(x) \end{aligned}$$

This means that $p(x)$ is increasing on the left side of $[c]$, and $p(x)$ is decreasing on the right side of $[c]$, guaranteeing that the maximum occurs at $[c]$. When c itself is an integer, then there are two values where the maximum occurs: both $c - 1$ and c . (Notation: $[c]$ means the integer part of c , that is, the greatest integer below c .)

Here we list the modes - without proofs - of the most important distributions. The proofs are excellent exercises for the reader.

1. Uniform distribution on $\{A, A+1, \dots, B-1, B\}$

Since all the possible values have the same probability, all are modes.

2. Hyper-geometrical distribution with parameters A, B, n

If $(n + 1) \frac{A+1}{A+B+2}$ is an integer, then there are two modes:

$$(n + 1) \frac{A + 1}{A + B + 2}$$

and

$$(n + 1) \frac{A + 1}{A + B + 2} - 1$$

If $(n + 1) \frac{A+1}{A+B+2}$ is not an integer, then there is only one mode:

$$\left\lfloor (n + 1) \frac{A + 1}{A + B + 2} \right\rfloor$$

3. Indicator distribution with parameter p

If $p < \frac{1}{2}$, then the mode is 0,

if $p > \frac{1}{2}$, then the mode is 1,

if $p = \frac{1}{2}$, then both values are the modes.

4. Binomial distribution with parameters n and p

If $(n + 1) p$ is an integer, then there are two modes:

$$(n + 1) p$$

and

$$(n + 1) p - 1$$

If $(n + 1) p$ is not an integer, then there is only one mode:

$$\lfloor (n + 1) p \rfloor$$

5. Poisson-distribution with parameter λ

If λ is an integer, then there are two modes:

$$\lambda$$

and

$$\lambda - 1$$

If λ is not an integer, then there is only one mode:

$$\lfloor \lambda \rfloor$$

6. Geometrical distribution (optimistic) with parameter p

The mode is 1.

7. Geometrical distribution (pessimistic) with parameter p

The mode is 0.

8. Negative binomial distribution (optimistic) with parameters r and p

If $\frac{r-1}{p}$ is an integer, then there are two modes:

$$\frac{r - 1}{p} + 1$$

and

$$\frac{r-1}{p}$$

If $\frac{r-1}{p}$ is not an integer, then there is only one mode:

$$\left\lfloor \frac{r-1}{p} \right\rfloor + 1$$

9. Negative binomial distribution (pessimistic) with parameters r and p

If $\frac{r-1}{p}$ is an integer, then there are two modes:

$$\frac{r-1}{p} - r + 1$$

and

$$\frac{r-1}{p} - r$$

If $\frac{r-1}{p}$ is not an integer, then there is only one mode:

$$\left\lfloor \frac{r-1}{p} \right\rfloor - r + 1$$

12 Expected value of discrete distributions

Formal definition of the expected value. Imagine a random variable X which has a finite or infinite number of possible values:

$$x_1, x_2, x_3, \dots$$

The probabilities of the possible values are denoted by

$$p_1, p_2, p_3, \dots$$

The possible values and their probabilities together constitute the distribution of the random variable. We may multiply each possible value by its probability, and we get the products:

$$x_1p_1, \quad x_2p_2, \quad x_3p_3, \quad \dots$$

Summarizing these products we get the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

If this series is absolutely convergent, that is

$$|x_1p_1| + |x_2p_2| + |x_3p_3| + \dots < \infty$$

then the value of the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

is a well defined finite number. As you learned in calculus this means that rearrangements of the terms of the series do not change the value of the series. Clearly, if all the possible values are greater than or equal to 0, then absolute convergence means simple convergence. If there are only a finite number of possible values, then absolute convergence is fulfilled. In case of absolute convergence, the value of the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

is called the **expected value** of the distribution or the expected value of the random variable X , and we say that the expected value **exists and it is finite**. The expected value is denoted by the letter μ or by the symbol $\mathbf{E}(X)$:

$$\mathbf{E}(X) = \mu = x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

Sigma-notation of a summation. The sum defining the expected value can be written like this:

$$\mathbf{E}(X) = \sum_i x_i p_i$$

or

$$\mathbf{E}(X) = \sum_x x p(x)$$

The summation, obviously, takes place for all possible values x of X .

Using Excel. In Excel, the function `SUMPRODUCT` can be used to calculate the expected value of X : if the x values constitute `array1` (a row or a column) and the $p(x)$ values constitute `array2` (another row or column), then

$$\text{SUMPRODUCT}(\text{array}_1; \text{array}_2)$$

is the sum of the products $xp(x)$, which is the expected value of X :

$$\mathbf{E}(X) = \sum_x x p(x) = \text{SUMPRODUCT}(\text{array}_1; \text{array}_2)$$

Mechanical meaning of the expected value: center of mass. If a point-mass distribution is considered on the real line, then - as it is known from mechanics - the center of mass is at the point:

$$\frac{x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots}{p_1 + p_2 + p_3 + \dots}$$

If the total mass is equal to one - and this is the case when we have a probability distribution - , then

$$p_1 + p_2 + p_3 + \dots = 1$$

and we get that the center of mass is at the point

$$\frac{x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots}{1} = x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots$$

which gives that the mechanical meaning of the expected value is the center of mass.

Law of large numbers. Now we shall derive the probabilistic meaning of the expected value. For this purpose imagine that we make N experiments for X . Let the experimental results be denoted by X_1, X_2, \dots, X_N . The average of the experimental results is

$$\frac{X_1 + X_2 + \dots + X_N}{N}$$

We shall show that if the expected value exist, and it is finite, then for large N , the average of the experimental results stabilizes around the expected value:

$$\frac{X_1 + X_2 + \dots + X_N}{N} \approx x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots = \mu = \mathbf{E}(X)$$

This fact, namely, that for a large number of experiments, the average of the experimental results approximates the expected value, is called the **law of large numbers** for the averages. In order to see that the law of large numbers holds, let the frequencies of the possible values be denoted by

$$N_1, N_2, N_3, \dots$$

Remember that

N_1 shows how many times x_1 occurs among X_1, X_2, \dots, X_N ,
 N_2 shows how many times x_2 occurs among X_1, X_2, \dots, X_N ,
 N_3 shows how many times x_3 occurs among X_1, X_2, \dots, X_N ,
and so on.

The relative frequencies are the proportions:

$$\frac{N_1}{N}, \frac{N_2}{N}, \frac{N_3}{N}, \dots$$

If N is large, then the relative frequencies stabilize around the probabilities:

$$\frac{N_1}{N} \approx p_1, \quad \frac{N_2}{N} \approx p_2, \quad \frac{N_3}{N} \approx p_3, \quad \dots$$

Obviously, the sum of all experimental results can be calculated so that x_1 is multiplied by N_1 , x_2 is multiplied by N_2 , x_3 is multiplied by N_3 , and so on, and then these products are added:

$$X_1 + X_2 + \dots + X_N = x_1 N_1 + x_2 N_2 + x_3 N_3 + \dots$$

This is why

$$\frac{X_1 + X_2 + \dots + X_N}{N} = \frac{x_1 N_1 + x_2 N_2 + x_3 N_3 + \dots}{N} = x_1 \frac{N_1}{N} + x_2 \frac{N_2}{N} + x_3 \frac{N_3}{N} + \dots$$

Since the relative frequencies on the right side of this equality, for large N , stabilize around the probabilities, we get that

$$\frac{X_1 + X_2 + \dots + X_N}{N} \approx x_1 p_1 + x_2 p_2 + x_3 p_3 + \dots = \mu = \mathbf{E}(X)$$

Remark. Sometimes it is advantageous to write the sum in the definition of the expected value like is:

$$\mathbf{E}(X) = \mu = \sum_x x p(x)$$

where the summation takes place for all possible values x .

The expected value may not exist! If the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

is not absolutely convergent, that is

$$|x_1p_1| + |x_2p_2| + |x_3p_3| + \dots = \infty$$

then one of the following 3 cases holds:

1. Either

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots = \infty$$

2. or

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots = -\infty$$

3. or the value of the series

$$x_1p_1 + x_2p_2 + x_3p_3 + \dots$$

is not well defined, because different rearrangements of the series may yield different values for the sum.

It can be shown that, in the first case, as N increases

$$\frac{X_1 + X_2 + \dots + X_N}{N}$$

will become larger and larger, and it approaches ∞ . This is why we may say that the expected exists, and its value is ∞ . In the second case, as N increases,

$$\frac{X_1 + X_2 + \dots + X_N}{N}$$

will become smaller and smaller, and it approaches $-\infty$. This is why we may say that the expected exists, and its value is $-\infty$. In the third case, as N increases,

$$\frac{X_1 + X_2 + \dots + X_N}{N}$$

does not approach to any finite or infinite value. In this case we say that the expected value does not exist.

In the following example, we give an example when the expected value is infinity, and thus, the sequence of averages goes to infinity.

Example 1. "What I pay doubles" - sequence of averages goes to infinity. We toss a coin until the first head (first "success"), and count how many tails ("failures") we get before the first head. If this number is T , then the amount of money I pay is

$X = 2^T$ forints. The amount of money is as much as if "we doubled the amount for each failure". We study the sequence of the money I pay.

In the following example, we give an example when the expected value does not exist, and thus, the sequence of averages does not converge.

Example 2. "What we pay to each other doubles" - averages do not stabilize. We toss a coin. If it is a head, then I pay a certain amount of money to my opponent. If it is a tail, then my opponent pays the certain amount of money to me. The amount of money is generated by tossing a coin until the first head (first "success"), and counting how many tails ("failures") we get before the first head. If this number is T , then the amount of money is $X = 2^T$ forints. The amount of money is as much as if "we doubled the amount for each failure". We study the sequence of the money I get, which is positive if my opponent pays to me, and it is negative if I pay to my opponent.

13 Expected values of the most important discrete distributions

Here we give a list of the formulas of the expected values of the most important discrete distributions. The proofs are given after the list. They are based mainly on algebraic identities and calculus rules. Some proofs would be easy exercises for the reader, others are trickier and more difficult.

1. **Uniform distribution on $\{A, A+1, \dots, B-1, B\}$**

$$\mathbf{E}(X) = \frac{A + B}{2}$$

2. **Hyper-geometrical distribution with parameters A, B, n**

$$\mathbf{E}(X) = n \frac{A}{A + B}$$

3. **Indicator distribution with parameter p**

$$\mathbf{E}(X) = p$$

4. **Binomial distribution with parameters n and p**

$$\mathbf{E}(X) = np$$

5. **Geometrical distribution (optimistic) with parameter p**

$$\mathbf{E}(X) = \frac{1}{p}$$

6. **Geometrical distribution (pessimistic) with parameter p**

$$\mathbf{E}(X) = \frac{1}{p} - 1$$

7. **Negative binomial distribution (optimistic) with parameters r and p**

$$\mathbf{E}(X) = \frac{r}{p}$$

8. **$\mathbf{E}(X)$ = Negative binomial distribution (pessimistic) with parameters r and p**

$$\mathbf{E}(X) = \frac{r}{p} - r$$

9. **Poisson-distribution with parameter λ**

$$\mathbf{E}(X) = \lambda$$

Proofs.

1. Uniform distribution on $\{A, A+1, \dots, B-1, B\}$

$$\begin{aligned} \mathbf{E}(X) &= \sum_x x p(x) = \sum_{x=A}^B x \frac{1}{B-A+1} = \\ &= \frac{1}{B-A+1} \sum_{x=A}^B x = \frac{1}{B-A+1} (B-A+1) \frac{A+B}{2} = \frac{A+B}{2} \end{aligned}$$

Since the distribution is symmetrical about $\frac{a+b}{2}$, it is natural that the expected value is $\frac{a+b}{2}$.

2. Hyper-geometrical distribution with parameters A, B, n

$$\begin{aligned} \mathbf{E}(X) &= \sum_x x p(x) = \\ &= \sum_{x=\max(0, n-B)}^{\min(n, A)} x \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} = \\ &= \sum_{x=\max(1, n-B)}^{\min(n, A)} x \frac{\binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} = \\ &= \sum_{x=\max(1, n-B)}^{\min(n, A)} \frac{x \binom{A}{x} \binom{B}{n-x}}{\binom{A+B}{n}} = \\ &= \sum_{x=\max(1, n-B)}^{\min(n, A)} \frac{A \binom{A-1}{x-1} \binom{B}{n-x}}{\frac{A+B}{n} \binom{A-1+B}{n-1}} = \\ &= n \frac{A}{A+B} \sum_{x=\max(1, n-B)}^{\min(n, A)} \frac{\binom{A-1}{x-1} \binom{B}{n-x}}{\binom{A-1+B}{n-1}} = \\ &= n \frac{A}{A+B} \sum_{x=\max(0, n-1-B)}^{\min(n-1, A-1)} \frac{\binom{A-1}{y} \binom{B}{n-1-y}}{\binom{A-1+B}{n-1}} = n \frac{A}{A+B} \end{aligned}$$

We replaced $x - 1$ by y , that is why we wrote $1 + y$ instead of x , and in the last step, we used that

$$\sum_{x=\max(0, n-1-B)}^{\min(n-1, A-1)} \frac{\binom{A-1}{y} \binom{B}{n-1-y}}{\binom{A-1+B}{n-1}} = 1$$

which follows from the fact that

$$\frac{\binom{A-1}{y} \binom{B}{n-1-y}}{\binom{A-1+B}{n-1}}$$

$$(\max(0, n-1-B) \leq x \leq \min(n-1, A-1))$$

is the weight function of the hyper-geometrical distribution with parameters $A-1, B, n-1$.

3. Indicator distribution with parameter p

$$\mathbf{E}(X) = \sum_x x p(x) = 0(1-p) + 1p = p$$

4. Binomial distribution with parameters n and p

$$\mathbf{E}(X) = \sum_x x p(x) =$$

$$\sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^n x \binom{n}{x} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^n n \binom{n-1}{x-1} p p^{x-1} (1-p)^{n-x} =$$

$$np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} =$$

$$np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} = np$$

We replaced $x - 1$ by y , that is why we wrote $1 + y$ instead of x , and in the last step, we used that

$$\sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} = 1$$

which follows from the fact that

$$\binom{n-1}{y} p^y (1-p)^{n-1-y} \quad \text{if } y = 0, 1, 2, \dots, n$$

is the weight function of the binomial distribution with parameters $n - 1$ and p .

5. **Geometrical distribution (optimistic) with parameter p .** We give two proofs. The first proof uses the techniques of summarizing geometrical series:

$$\begin{aligned} \mathbf{E}(X) &= \sum_x x p(x) = \\ & p + 2pq + 3pq^2 + 4pq^3 + \dots = \\ & \begin{array}{cccccc} p & + & pq & + & pq^2 & + & pq^3 & + & \dots \\ & + & pq & + & pq^2 & + & pq^3 & + & \dots \\ & & & + & pq^2 & + & pq^3 & + & \dots \\ & & & & & + & pq^3 & + & \dots \\ & & & & & & & \ddots & \\ & & & & & & & & = \end{array} \\ & \frac{p}{1-q} + \frac{pq}{1-q} + \frac{pq^2}{1-q} + \frac{pq^3}{1-q} + \dots = \\ & 1 + q + q^2 + q^3 + \dots = \\ & \frac{1}{1-q} = \frac{1}{p} \end{aligned}$$

The second proof uses the techniques of power series. Using the notation $q = 1 - p$, we get that

$$\begin{aligned} \mathbf{E}(X) &= \sum_x x p(x) = \\ & 1 p + 2 p(1-p) + 3 p(1-p)^2 + 4 p(1-p)^3 + \dots = \\ & 1 p + 2 pq + 3 pq^2 + 4 pq^3 + \dots = \end{aligned}$$

$$p (1 + 2q + 3q^2 + 4q^3 + \dots) = p \frac{1}{(1-q)^2} = p \frac{1}{p^2} = \frac{1}{p}$$

We used the identity

$$1 + 2q + 3q^2 + 4q^3 + \dots = \frac{1}{(1-q)^2}$$

which is proved by first considering the given infinite series as the derivative of a geometrical series, then taking the closed form of the geometrical series, and then differentiating the closed form:

$$\begin{aligned} 1 + 2q + 3q^2 + 4q^3 + \dots &= \\ \frac{d}{dq} (1 + q + q^2 + q^3 + q^4 \dots) &= \\ \frac{d}{dq} \left(\frac{1}{1-q} \right) &= \frac{d}{dq} ((1-q)^{-1}) = (1-q)^{-2} = \frac{1}{(1-q)^2} \end{aligned}$$

6. **Geometrical distribution (pessimistic) with parameter p .** Since the pessimistic geometrical distribution can be derived from the optimistic by a shift of 1 unit to the left, the expected value of the pessimistic geometrical distribution is equal to the expected value of the optimistic geometrical distribution minus 1.

$$\mathbf{E}(X) = \frac{1}{p} - 1$$

7. **Negative binomial distribution (optimistic) with parameters r and p**

$$\begin{aligned} \mathbf{E}(X) &= \sum_x x p(x) = \\ \sum_{x=r}^{\infty} x \binom{x-1}{r-1} p^r (1-p)^{x-r} &= \\ \sum_{x=r}^{\infty} r \binom{x}{r} \frac{p^{r+1}}{p} (1-p)^{x-r} &= \\ \frac{r}{p} \sum_{x=r}^{\infty} \binom{x}{r} p^{r+1} (1-p)^{x-r} &= \\ \frac{r}{p} \sum_{x=r+1}^{\infty} \binom{x-1}{r} p^{r+1} (1-p)^{x-1-r} &= \\ \frac{r}{p} \sum_{x=r+1}^{\infty} \binom{x-1}{r} p^{1+r} (1-p)^{x-1-r} &= 1 \end{aligned}$$

which follows from the fact that

$$\binom{y-1}{r} p^{1+r} (1-p)^{y-1-r} \quad \text{if } y = r+1, r+2, r+3, \dots$$

is the weight function of the (optimistic) negative binomial distribution with parameters $r+1$ and p .

8. **Negative binomial distribution (pessimistic) with parameters r and p .** Since the pessimistic negative binomial distribution can be derived from the optimistic by a shift of r units to the left, the expected value of the pessimistic negative binomial distribution is equal to the expected value of the optimistic negative binomial distribution minus r .

$$\mathbf{E}(X) = \frac{r}{p} - r$$

9. **Poisson-distribution with parameter λ**

$$\mathbf{E}(X) = \sum_x x p(x) =$$

$$\sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} =$$

$$\sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} =$$

$$\sum_{x=1}^{\infty} \lambda \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} =$$

$$\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} =$$

$$\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} = \lambda$$

We replaced $x-1$ by y , that is why we wrote $1+y$ instead of x , and in the last step, we used that

$$\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} = 1$$

which follows from the fact that

$$\frac{\lambda^y}{y!} e^{-\lambda} \quad (y = 0, 1, 2, \dots)$$

is the weight function of the Poisson distribution with parameter λ .

Example. Shooting stars. If you watch the sky from a peak of "Kékes-tető" (highest peak in Hungary) around midnight in August, you will see shooting-stars. Assume that the amount of time between two shooting-stars, which is a random quantity, is 10 minutes on the average. If we watch the sky for 15 minutes, then how much is the probability that we see exactly 2 shooting-stars?

Remark. It is not difficult to figure out wrong arguments to get the wrong answers: 0.5 or 0.75. The reader will enjoy to find out these wrong arguments.

Solution. The number of shooting-stars visible during 15 minutes is a random variable. The following facts are obvious:

1. the number of meteors in the space is vary large, and
2. for each meteor the probability that it causes a shooting-star during our 10 minute is small, and
3. the meteors cause a shooting-stars independently of each other,

These facts guarantee that the number of shooting-stars visible during 15 minutes follows a Poisson distribution. The parameter λ of this distribution is equal to the expected value of the number of shooting-stars visible during 15 minutes. Since the amount of time between two shooting-stars is 10 minutes in the average, the average number of shooting-stars visible during 15 minutes is 1.5. Thus, $\lambda = 1.5$. This is why the answer to the question is

$$\begin{aligned} \mathbf{P}(\text{We see exactly 2 shooting-stars}) &= \frac{\lambda^2}{2!} e^{-\lambda} = \\ &= \text{POISSON}(2; 1,5; \text{FALSE}) \approx 0,251 \end{aligned}$$

14 Expected value of a function of a discrete random variable

When a random variable X is considered, and $y = t(x)$ is a function, then $Y = t(X)$ clearly defines another random variable. The random variable Y is called the **function of the random variable** X . If we make N experiments for the random variable X , and we substitute the experimental results X_1, X_2, \dots, X_N into the function $y = t(x)$, we get the values $t(X_1), t(X_2), \dots, t(X_N)$. Their average is

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N}$$

It can be shown that - under some conditions - if N is large, then this average also stabilizes around a non-random value. We show this fact. Obviously,

$$t(X_1) + t(X_2) + \dots + t(X_N) = t(x_1)N_1 + t(x_2)N_2 + t(x_3)N_3 + \dots$$

This is why

$$\begin{aligned} \frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N} &= \frac{t(x_1)N_1 + t(x_2)N_2 + t(x_3)N_3 + \dots}{N} = \\ &= t(x_1)\frac{N_1}{N} + t(x_2)\frac{N_2}{N} + t(x_3)\frac{N_3}{N} + \dots \end{aligned}$$

Since the relative frequencies in this formula, for large N , stabilize around the probabilities, we get that

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N} \approx t(x_1)p_1 + t(x_2)p_2 + t(x_3)p_3 + \dots$$

The non-random value on the right side of this formula is the expected value of $t(X)$:

$$\mathbf{E}(t(X)) = t(x_1)p_1 + t(x_2)p_2 + t(x_3)p_3 + \dots$$

Sometimes it is advantageous to write the sum in the following form:

$$\mathbf{E}(t(X)) = \sum_i t(x_i) p_i$$

or

$$\mathbf{E}(t(X)) = \sum_x t(x) p(x)$$

where the summation takes place for all possible values x . We emphasize again that if N is large, then the average of the values $t(X_1), t(X_2), \dots, t(X_N)$ is close to the expected value of $t(X)$:

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N} \approx \mathbf{E}(t(X))$$

The condition for the existence and finiteness of the expected value of $t(X)$ is that the series

$$t(x_1) p_1 + t(x_2) p_2 + t(x_3) p_3 + \dots$$

is absolutely convergent, which means that

$$|t(x_1)| p_1 + |t(x_2)| p_2 + |t(x_3)| p_3 + \dots < \infty$$

The expected value may not exist! If the infinite sum

$$t(x_1) p_1 + t(x_2) p_2 + t(x_3) p_3 + \dots$$

is not absolute convergent, then its value is either ∞ or $-\infty$ or its value is not well defined, because different rearrangements of the series may yield different values for the sum. In the first case, it can be proven that as N increases,

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N}$$

will become larger and larger, and it approaches ∞ . This is why we may say that the expected value of $t(X)$ is ∞ . In the second case, it can be proven that as N increases,

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N}$$

will become smaller and smaller, and it approaches $-\infty$. This is why we may say that the expected value of $t(X)$ is $-\infty$. In the third case, it can be proven that as N increases,

$$\frac{t(X_1) + t(X_2) + \dots + t(X_N)}{N}$$

does not approach to any finite or infinite value. In this case we say that the expected value of $t(X)$ does not exist.

Using Excel. In Excel, the function `SUMPRODUCT` can be used to calculate the expected value of $t(X)$: if the x values constitute `array1` (a row or a column) and the $p(x)$ values constitute `array2` (another row or column) and the $t(x)$ values constitute `array3` (a third row or column), then

$$\text{SUMPRODUCT}(\text{array}_3; \text{array}_2)$$

is the sum of the products $t(x)p(x)$, which is the expected value of $t(X)$:

$$\mathbf{E}(t(X)) = \sum_x t(x) p(x) = \text{SUMPRODUCT}(\text{array}_3; \text{array}_2)$$

15 Moments of a discrete random variable

The expected value of X^n is called the n th moment of X :

$$\mathbf{E}(X^n) = x_1^n p_1 + x_2^n p_2 + x_3^n p_3 + \dots$$

or, using the other notations:

$$\mathbf{E}(X^n) = \sum_x x^n p(x)$$

The first moment coincides with the expected value. Among all moments, the first and the second moment play the most important role. For emphases, we repeat the definition of the second moment: the expected value of X^2 is called the **second moment** of X :

$$\mathbf{E}(X^2) = x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 + \dots$$

or, using the other notations:

$$\mathbf{E}(X^2) = \sum_x x^2 p(x)$$

Using Excel. In Excel, the function `SUMPRODUCT` can be used to calculate a moment of X : if the x values constitute `array1` (a row or a column) and the $p(x)$ values constitute `array2` (another row or column) and the n th powers of the x values constitute `array3` (a third row or column), then

$$\text{SUMPRODUCT}(\text{array}_3; \text{array}_2)$$

is the sum of the products $x^n p(x)$, which is the n th moment of X :

$$\mathbf{E}(X^n) = \sum_x x^n p(x) = \text{SUMPRODUCT}(\text{array}_3; \text{array}_2)$$

Using Excel. In Excel, the second moment of X can be calculated also like this: if the x values constitute `array1` (a row or a column) and the $p(x)$ values constitute `array2` (another row or column), then

$$\text{SUMPRODUCT}(\text{array}_1; \text{array}_1; \text{array}_2)$$

is the sum of the products $x x p(x)$, which is the second moment of X :

$$\mathbf{E}(X^2) = \sum_x x^2 p(x) = \sum_x x x p(x) = \text{SUMPRODUCT}(\text{array}_1; \text{array}_1; \text{array}_2)$$

As an example, we calculate here the second moment of the binomial distribution.

Second moment of the binomial distribution with parameters n and p

$$np + n^2p^2 - np^2$$

Proof.

$$\begin{aligned}\mathbf{E}(X^2) &= \sum_x x^2 p(x) = \\ &= \sum_{x=0}^n x^2 \binom{n}{x} p^x (1-p)^{n-x} = \\ &= \sum_{x=1}^n x^2 \binom{n}{x} p^x (1-p)^{n-x} = \\ &= \sum_{x=1}^n n x \binom{n-1}{x-1} p^x (1-p)^{n-x} = \\ &= np \sum_{x=1}^n x \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} =\end{aligned}$$

Now we replace $x - 1$ by y , that is we write $1 + y$ instead of x , and we get:

$$np \sum_{y=0}^{n-1} (1+y) \binom{n-1}{y} p^y (1-p)^{n-1-y} =$$

Now the sum splits into the sum of two sums:

$$np \left[\left(\sum_{y=0}^{n-1} 1 \binom{n-1}{y} p^y (1-p)^{n-1-y} \right) + \left(\sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{n-1-y} \right) \right] =$$

Here, inside the bracket, the first sum is the sum of the terms of the binomial distribution with parameters $n - 1$ and p , so it is equal to 1. The second sum is the expected value of the binomial distribution with parameters $n - 1$ and p , so it is equal to $(n - 1)p$. This is why we get:

$$np [1 + (n - 1)p] = np + n^2p^2 - np^2$$

16 Average distance from the average, variance, standard deviation, etc.

Both the median and the average define a kind of "center" for a data-set, or for a distribution. It is important to have some characteristics to measure the deviation from the center for a data-set and for a distribution. In this chapter, we shall learn such characteristics.

If z_1, z_2, \dots, z_N is a data-set, consisting of numbers, then their average is a well-known characteristic of the data-set, which will be denoted by \bar{z}_N or, for simplicity, by \bar{z} :

$$\bar{z}_N = \bar{z} = \frac{z_1 + z_2 + \dots + z_N}{N}$$

The average shows where the center of the data-set is.

It is important for us to know how far the data are from the average. This is why we consider the distance (that is, the absolute value of the difference) between the data elements and the average:

$$|z_1 - \bar{z}|, |z_2 - \bar{z}|, \dots, |z_N - \bar{z}|$$

The average of these distances is a characteristic of how far the data elements, in the average, are from their average:

$$\frac{|z_1 - \bar{z}| + |z_2 - \bar{z}| + \dots + |z_N - \bar{z}|}{N}$$

This quantity is called the **average distance from the average**.

Using Excel. In Excel, for a data-set, the function AVEDEV calculates the average distance from the average.

If, before taking the average, instead of the absolute value, we take the square of each difference, we get another characteristic of the data-set, the average squared distance from the average, which is called the **variance** of the data-set:

$$\frac{(z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + \dots + (z_N - \bar{z})^2}{N}$$

The square root of the variance is called the **standard deviation** of the data-set:

$$\sqrt{\frac{(z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + \dots + (z_N - \bar{z})^2}{N}}$$

Using a calculator. Most calculators have a key to determine not only the average of a data-set, but the average distance from the average, the variance and the standard deviation, as well.

Using Excel. In Excel, for a data-set, the function `VARP` calculates the variance, and the function `STDEV` calculates the standard deviation.

Sample variance and sample standard deviation in Excel. In Excel, the functions `VAR` and `STDEV` calculate the so called **sample variance** and **sample standard deviation**. The sample variance and sample standard deviation are defined almost the same way as the variance and standard deviation, but the denominator is $N - 1$ instead of N :

$$\frac{(z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + \dots + (z_N - \bar{z})^2}{N - 1}$$

$$\sqrt{\frac{(z_1 - \bar{z})^2 + (z_2 - \bar{z})^2 + \dots + (z_N - \bar{z})^2}{N - 1}}$$

The advantage of taking $N - 1$ instead of N becomes clear in statistics. We will not use the functions `VAR` and `STDEV`.

Recall that if we make a large number of experiments for a random variable X , then the average of the experimental results, in most cases, stabilizes around a non-random value, the expected value of the random variable, which we denote by μ :

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} \approx$$

$$\mu = \begin{cases} \sum xp(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} x p(x) & \text{in the continuous case} \end{cases}$$

The average distance from the average of the experimental results, long sequence of experiments, also stabilizes around a non-random value, which we call the **average distance from the average** of the random variable or of the distribution, which we denote by d :

$$\frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + \dots + |X_N - \bar{X}|}{N} \approx$$

$$\frac{|X_1 - \mu| + |X_2 - \mu| + \dots + |X_N - \mu|}{N} \approx$$

$$d = \begin{cases} \sum |x - \mu| p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} |x - \mu| f(x) dx & \text{in the continuous case} \end{cases}$$

The variance of the experimental results, in a long sequence of experiments, also stabilizes around a non-random value, which we call the **variance** of the random variable or of the distribution, which we denote by σ^2 :

$$\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_N - \bar{X})^2}{N} \approx$$

$$\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N} \approx$$

$$\mathbf{VAR}(X) = \sigma^2 = \begin{cases} \sum (x - \mu)^2 p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{in the continuous case} \end{cases}$$

The standard deviation of the experimental results, which is the square root of the variance, in a long sequence of experiments, obviously stabilizes around the square root of σ^2 , that is, σ , which we call the **standard deviation** of the random variable or of the distribution, which we denote by σ :

$$\sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_N - \bar{X})^2}{N}} \approx$$

$$\sqrt{\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N}} \approx$$

$$\mathbf{SD}(X) = \sigma = \begin{cases} \sqrt{\sum (x - \mu)^2 p(x)} & \text{in the discrete case} \\ \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx} & \text{in the continuous case} \end{cases}$$

These non-random values are characteristics of the random variable and of the distribution. Among these three characteristics the variance and the standard deviation play a much more important theoretical and practical role than the average distance from the average.

Mechanical meaning of the variance. The mechanical meaning of the variance is the inertia about the center, because it is calculated by the same formula as the variance:

$$\begin{cases} \sum (x - \mu)^2 p(x) & \text{in the discrete case} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{in the continuous case} \end{cases}$$

Remark. It may seem a little bit strange that the notion of the variance and the standard deviation play a more important role than the notion of the average distance from the average. The reason is that the variance and the standard deviation satisfy a rule which is very important both for the theory and the practice. Namely, it is true that the variance of the sum of independent random variables equals the sum of the variances of the random variables, or equivalently the standard deviation of the sum of independent random variables equals to the sum of the squares of the standard deviations of the random variables. Such a general rule does not hold for the average distance from the average.

The variance is very often calculated on the basis of the following relation.

The variance equals the second moment minus the expected value squared:

$$\sum_x (x - \mu)^2 p(x) = \sum_x x^2 p(x) - \mu^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

in the continuous case.

The proof of these relations is quite simple. In the continuous case:

$$\begin{aligned} \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx &= \\ \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x) dx &= \\ \int_{-\infty}^{\infty} x^2 f(x) dx - \int_{-\infty}^{\infty} 2x\mu f(x) dx + \int_{-\infty}^{\infty} \mu^2 f(x) dx &= \\ \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx &= \\ \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \mu + \mu^2 1 &= \\ \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 & \end{aligned}$$

In the discrete case, the integral is replaced by summation, $f(x)$ is replaced by $p(x)$.

Using Excel. In Excel, the variance of a discrete distribution given by numerical values can be calculated like this: if the distribution is arranged in a table-form so that the x values constitute `array1` (a row or a column) and the associated $p(x)$ values constitute `array2` (another row or column) then we may calculate the expected value μ by the

$$\mu = \text{SUMPRODUCT}(\text{array}_1; \text{array}_2)$$

command, and then we may calculate $(x - \mu)^2$ for each x , and arrange these squared distances into `array3`. Then the variance is

$$\sigma^2 = \text{SUMPRODUCT}(\text{array}_3; \text{array}_2)$$

and the standard deviation is

$$\sigma = \text{SQRT}(\text{SUMPRODUCT}(\text{array}_3; \text{array}_2))$$

Steiner's equality. The second moment of a distribution about a point c is equal to the variance plus the difference between the expected value and c squared:

$$\sum_x (x - c)^2 p(x) = \sum_x (x - \mu)^2 p(x) + (\mu - c)^2 = \sigma^2 + (\mu - c)^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x - c)^2 f(x) dx = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx + (\mu - c)^2 = \sigma^2 + (\mu - c)^2$$

in the continuous case.

Steiner's inequality. The second moment of a distribution about any point c is greater than the variance, and equality holds only if $c = \mu$. In other words, the second moment of a distribution about any point c is minimal, if $c = \mu$, and the minimal value is σ^2 :

$$\sum_x (x - c)^2 p(x) \geq \sigma^2$$

in the discrete case, and

$$\int_{-\infty}^{\infty} (x - c)^2 f(x) dx \geq \sigma^2$$

in the continuous case. Equality holds if and only if $c = \mu$.

The proof of Steiner's equality, for the continuous case:

$$\begin{aligned} & \int_{-\infty}^{\infty} (x - c)^2 f(x) dx = \\ & \int_{-\infty}^{\infty} ((x - \mu) + (\mu - c))^2 f(x) dx = \\ & \int_{-\infty}^{\infty} ((x - \mu)^2 + 2(x - \mu)(\mu - c) + (\mu - c)^2) f(x) dx = \\ & \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx + \int_{-\infty}^{\infty} 2(x - \mu)(\mu - c)f(x) dx + \int_{-\infty}^{\infty} (\mu - c)^2 f(x) dx = \\ & \sigma^2 + 2(\mu - c) \int_{-\infty}^{\infty} (x - \mu)f(x) dx + (\mu - c)^2 \int_{-\infty}^{\infty} f(x) dx = \\ & \sigma^2 + 2(\mu - c) 0 + (\mu - c)^2 1 = \\ & \sigma^2 + 0 + (\mu - c)^2 = \\ & \sigma^2 + (\mu - c)^2 \end{aligned}$$

In the above argument, we used the fact that

$$\begin{aligned} \int_{-\infty}^{\infty} (x - \mu)f(x) dx &= \\ \int_{-\infty}^{\infty} xf(x) dx - \int_{-\infty}^{\infty} \mu f(x) dx &= \\ \int_{-\infty}^{\infty} xf(x) dx - \mu \int_{-\infty}^{\infty} f(x) dx &= \\ \mu - \mu \cdot 1 &= 0 \end{aligned}$$

In the discrete case, the integral is replaced by summation, $f(x)$ is replaced by $p(x)$. The Steiner's inequality is an obvious consequence of the Steiner's equality.

Steiner's equality in mechanics. Steiner's equality in mechanics is well-known: the inertia about a point c is equal to the inertia about the center of mass plus inertia about the point c as if the total amount of mass were in the center of mass.

Steiner's inequality in mechanics. Steiner's inequality in mechanics is well-known: the inertia about a point c which is different from the center of mass is greater than the inertia about the center of mass.

Variance and standard deviation of some distributions:

1. Binomial distribution.

The second moment of the binomial distribution is

$$\mathbf{E}(X^2) = n^2p^2 - np^2 + np$$

The expected value of the binomial distribution is

$$\mathbf{E}(X) = np$$

So, the variance is

$$\mathbf{VAR}(X) = (n^2p^2 - np^2 + np) - (np)^2 = np - np^2 = np(1 - p)$$

Thus, the standard deviation of the binomial distribution is

$$\mathbf{SD} = \sqrt{np(1 - p)}$$

2. Uniform distribution

The second moment of the uniform distribution on an interval $(A; B)$ is

$$\frac{A^2 + AB + B^2}{3}$$

The expected value of the uniform distribution is

$$\frac{A + B}{2}$$

So, the variance is

$$\mathbf{VAR}(X) = \left(\frac{A^2 + AB + B^2}{3} \right) - \left(\frac{A + B}{2} \right)^2 = \frac{(B - A)^2}{12}$$

Thus, the standard deviation of the uniform distribution is

$$\mathbf{SD} = \frac{(B - A)}{\sqrt{12}}$$

3. Exponential distribution

The second moment of the exponential distribution is

$$\mathbf{E}(X^2) = \frac{2}{\lambda^2}$$

The expected value of the exponential distribution is

$$\mathbf{E}(X) = \frac{1}{\lambda}$$

So, the variance is

$$\mathbf{VAR}(X) = \left(\frac{2}{\lambda^2} \right) - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}$$

Thus, the standard deviation of the exponential distribution is

$$\mathbf{SD} = \frac{1}{\lambda}$$