Classical probability

This chapter starts with an outline of how the notion ‘a probability’ was developed. The first scientific approach to it is now called ‘a classical probability’, as well as ‘a Laplace probability’. This probability is still correctly working when we deal with finite sets. It underwent changes – via geometrical probability and von Mises probability – to transform into so-called ‘an axiomatic probability’, or ‘a Kolmogorov probability’, or simply ‘a probability’. We note that the classical probability is a map satisfying some rules (and, next, some of these rule will serve as defining ones to the Kolmogorov probability).
Next, we familiarize with such notions as ‘a conditional probability’ and such formulas as ‘Bayes’ rule’. The last mentioned makes possible to treat the probability in a Bayesian way; this is, the probability can be seen in terms of updating the belief in a hypothesis when new data arrive. We familiarize this approach (aka a Bayesian inferrence) via examples concerning the quality of medical tests and the quality of the quality control (examining the quality of products made in a factory). In particular, we meet such terms as ‘a priori probability’, ‘a posteriori probability’, ‘a cross tab matrix’, ‘a cross frequency matrix’ and ‘an error matrix’, ‘a type I error’ and ‘a type II error’.

Very short history of the probability

The Latin verb *probare* means to prove worthly, to try, to test, to esteem, to make credible, to show, to judge by trial. In 14th century it was assimilated into the English language in two variations: as *to prove* and as *to probe*. In Latin the gerund formed from *probare* is *proba* (a test, an examination), in English the gerunds formed from *to prove* and *to probe* are a *proof* (the act of proving, when there is established the truth, the validity, the authenticity) and a *probe* (the act of probing, i.e., the act of examination with a probe).

The Latin word *probitas* means a probity (= complete and confirmed integrity), a honesty (≈ truthfulness, sincerity), an uprightness (≈ righteousness; rightnous = morally upright; without guilt or sin; in accordance with virtue or morality). In Europe legal institutions expected a witness to present his/her probity, to be honest, rightnous, honorable (*probus*), sincere (*sincerus*), true (*verus*). This expectation was often correlated with the witness's nobility.

Another Latin gerund formed form the verb *probare* is *probabilitas* – a credibility (*fides*), a possibility (*facultas*), an expectance (*spes*), a likelihood (*verisimilitude*), a probability (*veri similitudo* – true likeness, similarity, resemblance). All the meanings of *to probe* and both nouns and adjectives (such as provable = capable of being demonstrated or proved, and probable = likely to happen or to be true) formed from *to probe* compose what is of the top importance in this area of the human activity which is called theory of probability and statistics. It is widely accepted that it became to be a science in 1813, when Jacob Bernoulli's *Ars conjectandi* was published, and took its first complete form in 1812, when Pierre Simon Laplace in his *Théorie analytique des probabilities* (2nd edition) clearly postulated the definition what a probability is. Earlier to this date people used the notion 'probability' and 'probable' without explicit definitions. Obviously, in the years passing by they wanted (and, because of observed cases) they needed to precise what these notions mean.

Thanks to contributions given in 1654 by Pierre de Fermat and Blaise Pascal, in 1657 by Christiaan Huygens, by Jacob Bernoulli (1813), in 1718 by Abraham de Moivre, in 1812 Laplace worked out, in its final form, so-called classical definition of the probability. It underwent further investigations – let's mention here proposals given by Georges Leclerc Marquis de Buffon in 1733 and Richard von Misses in 1931. The search for the definition of the probability was successfully finalized with so-called axiomatic definition – it was given in 1933 by Andriey Kolmogorov.

We will familiarize Kolmogorov's definition later on. In this chapter (when we deal with finite sets only) we will apply classical definition of the probability – and this is good enough to not commit serious errors in the subject.

Classical probability

We first rather tell, what the notion 'a (classical) probability' stands for, its definition is given later on. It uses the terms

'an elementary event' [1)]
'an equiprobable event' [2)]
and 'independent events' [3)],

but there is taken no care what do these three notions mean and starts with a given finite set Ω of all possible elementary events. This set is called a sample space. An arbitrary subset A of Ω is called an event, and the probability that A occurs is the quotient of two cardinalities: that of all events favorable to A and that of the set Ω; in symbols it can be written down as follows:

$$\Pr(A) := \frac{|A|}{|\Omega|},$$

where $|A|$ – the number of events which favor the event A,
$|\Omega|$ – the number of elementary events.

This definition can be applied to situations when only finite cardinalities are involved, when Ω is a finite set. In the flow of years it appeared that the

---

[1)] aka an **atom event**. In an analogous way as in the chemistry when (to 1897 when Joseph John Thomson discovered electrons, subatomic particles, and showed they are in every atom) an elementary, indivisible, the smallest component of a matter of the universe was an atom, one can think of an atom event as of an event which is not composed of any other events. This approach to what an atom event is was saved in Kolmogorov's definition.

[2)] two events are equiprobable is they are of the same probability; it makes that in the classical definition of the probability there is reference to the definiendum (what it being defined), so it is *idem per ipsum* reference, aka a circular reference, a self-reference. The self-reference is the consequence of Aristotle's law of non-contradiction. The **law of contradiction** ($\sim\{p \,\&\, \sim p\}$: two propositions p and not p are mutually exclusive, contradictory statements contradictory statements cannot both be true in the same sense at the same time) and two other ones: the **law of identity** ($p = p$: p is p, and it is not not p) and the **law of excluded middle** ($p \vee \sim p$: for any statement p, there is true either p or its negation $\sim p$) are three basic principles of classic laws of thought, they are fundamental philosophical axioms (and without them the logic and the mathematics become impossible). In mathematics the self-reference is not allowed in definitions (and to avoid the impossibility to describe some objects and rules we assume so-called primitive objects and axioms), but it is quite natural to discuss recurrences.

[3)] events are understood to be independent if any of them impacts the other that they are independent. In the set-theory approach the (in)dependence is understood as follows: two events A and B are independent if the event $A \cap B$ is impossible, this is $A \cap B = \varnothing$.

considerations on probability embraces also cases with infinite Ω; in more details: the cardinality of the set of elementary events can be $\aleph_0$ (aleph zero, the cardinality of the set ***N*** of natural numbers) and *c* (continuum, the cardinality of the set ***R*** of real numbers) and that the dependency of events is much more complicated matter than it seems to be in Laplace's era. Searching for the definition of the probability which deals with both situations gave birth to what is now called a geometrical probability and a von Misses' limit, and were closed by A.Kolmogorov in 1933 with his axioms – all these three ideas will be presented later on.

After the above introduction let's introduce the notion 'probability' in formal way. We start with saying that in the theory of probability and statistics an **experiment** is any controlled, repeatable process.

Examples of experiments are:

a) tossing a coin,
b) rolling a die [4)],
c) taking blindly a ball from a bag,
d) random selecting a name out of the list.

Experiments can be composed of multiple actions. Examples of such experiments are:

e) throwing a coin twice,
f) simultaneously flipping a coin and casting a die,
g) noting, at every full hour of the day, how many cars are in the parking plot,
h) noting how many customers stay in the shop at every day of the month.

A single, specific possible result of an experiment is called an **outcome**. The set of all possible outcomes is called a **sample space** (in this experiment). Usually, the sample space is denoted by Ω.

There are no other outcomes in the experiment a) than the head (H) and the tail (T), so Ω = { H, T }. If we identify H and 1, and T and 0, then all possible outcomes are the numbers 1 and 0, so Ω = { 0, 1 }.

When throwing a die, every outcome is the number of points exhibited on (or showed by) a die, so Ω = { 1, 2, 3, 4, 5, 6 }.

[4)] In English the word 'a die' has its plural form 'dice'. In USA the word 'dice' stands for both singular and plural form. The word ' a die' was assimilated to English via the French *dé* from Latin *datum* (something which is given or played). When we speak of 'a die (or a dice)' we think of a cube (French *carré*) with each of its six faces showing a different number of dots (aka pips) from 1 to 6 and we assume it is fair, unbiased, not deceived, not cheated, not biased.

If a bag contains seven balls and every one of them is in one of Newtonian rainbow colors [5] (they are: red, orange, yellow, green, blue, indigo and violet) and the experiment consists in pulling one ball, then

$$\Omega = \{ \mathrm{R}, \mathrm{O}, \mathrm{Y}, \mathrm{G}, \mathrm{B}, \mathrm{I}, \mathrm{V} \},$$

where R stands for the event 'red ball is drawn out', O denotes the event 'orange ball is chosen' etc.

Instead of letter designations for colors one can identify them by their wavelengths, e.g., the length of the red light is about 650 nm, that of violet is about 420 nm (1 nanometer is 10^{-9} m). Then the sample space appears as

$$\Omega = \{ 700, 620, 580, 530, 470, 445, 420 \}.$$

If colors are identified with their frequencies, e.g., the frequency of the yellow light is about 580 THz (1 terahertz is 10^{12} hertz, 1 Hz = 1/(1 s), s stands for the second), then the same sample space is, up to some accuracy,

$$\Omega = \{ 428, 484, 517, 566, 638, \ldots, 714 \}.$$

Any subset of the sample space Ω is called an **event**. As Ω is a set, for every event there exists it complement (denoted as A' or A^{C}), and every event is the sum of some elements. In the next let's denote the set of all events by S.

Notice that always Ω includes itself (Ω) and the empty set ($\varnothing$); these two events are referred to as a **sure event** and an **impossible event**, resp. We will below see that, in the probability theory (as well as in practice), these both improper subsets of the sample space play special roles:

a sure event is the event that always happens, that surely takes place,
an impossible event is the event which never occurs.

Notice also, please, that in Ω there are always A and the **opposite event** to A; it is denoted as A^{C} and it is, in the set-theoretical terms, the **complement** to A,

$$A^{\mathrm{C}} := \Omega \setminus A.$$

[5] The phenomenon that a glass prism disassembles the white light was first noted by Isaac Newton and described in his book *Opticks or treatise of the reflexions, refractions, inflexions and colours of light* published in 1704. Earlier, in 1671, he called the collections of colors obtained in this way as a spectrum, and he distinguished seven colors – they all form what is called 'Newtonian rainbow palette'. The modern meaning of original Newtonian colors is changed: his blue corresponds to what is today called cyan, and his indigo is today referred to as blue.

Now we are going to define the term ‘probability’ in the way good enough it to work when the sample space is finite. Operationally, the probability of an event A is a measure of the likelihood that A occurs, a measure that the event A takes place, that A holds. The **probability** (that) the event A occurs is denoted by $\Pr(A)$ and defined via the formula

$$\Pr(A) := \frac{|A|}{|\Omega|},$$

where $|A|$ and $|\Omega|$ are cardinalities of A and of Ω, resp.
It says that the probability $\Pr(A)$ is the quotient,

whose numerator is the number of outcomes belonging to A,
and the denominator is the number of outcomes in the sample space.

This can also be expressed as the quotient,

whose numerator is the number of events realizing the event A
(one can say: the number of events favoring the event A)
and the denominator is the number of all possible events.

In the form cited above this definition was first formulated by Pierre-Simon Laplace and that’s why it is also known as a **Laplace probability**. P.-S.Laplace clarified what was conceived earlier, and this is remembered via another name for this probability: a **classical probability**. Let’s emphasize: in theoretical-set terminology (where, as we know, the events are simply sets) the Laplace probability of an event A (being a sum of elements of Ω) is the quotient of two cardinalities: that of A and that of Ω.

Example (2 balls pulled out of 3 red and 4 green ones). In a bag there are 7 balls, 3 of them are red and other ones are green. At the same time we draw out two balls. We want to know which is the probability these two balls are green.
Although balls are identical, let's distinguish them by labeling them as r, s, t (three red balls) and g, h, i, j (four green balls). Balls are taken off at the same time, so the order is no important, and all outcomes are (instead of $\{a, b\}$ we write ab):

balls in different colors: rg, rh, ri, rj,
sg, sh, si, sj,
tg, th, ti, tj,
both red balls: rs, rt, st,
both green balls: gh, gi, gj, hi, hj, ij.

There is no other outcome of the experiment, Ω = 'there are blindly selected two balls out of 3 red ones and 4 green ones'. These 21 pairs form the sample space Ω, in consequence $|\Omega| = 21$.
The asked event is A = 'both balls are green'. We see A = { gh, gi, gj, hi, hj, ij }, so $|A| = 6$.
Therefore

$$\Pr(A) = \frac{|A|}{|\Omega|} = \frac{6}{21} = \frac{2}{7}.$$

The same result can be produced without explicit listing of outcomes. To do it we directly apply combinatorics:

a) the number of ways to pull 2 balls out off 7 balls is $\binom{7}{2} = \frac{7 \cdot 6}{2 \cdot 1} = 21$,

b) there are $\binom{4}{2} \cdot \binom{3}{0} = \frac{4 \cdot 3}{2 \cdot 1} \cdot 1 = 6$

ways to realize the event A = 'Draw two green balls from the box containing 3 red balls and 4 green balls' (really: these 2 green balls have to be taken out off all 4 green balls, and no ball can be drawn out of 3 red balls).

The quotient of both numbers (6 ways favoring the event A against 21 possible ways) gives

$$\frac{\binom{4}{2} \cdot \binom{3}{0}}{\binom{7}{2}} = \frac{6}{21} = \frac{2}{7}.$$

Example ended.

One can check that the function Pr which is defined on the set S of all event so that

a) it assumes values in the interval <0, 1>: i.e., for any event A, $0 \le \Pr(A) \le 1$,
b) it equals 0 for the impossible event: $\Pr(\varnothing) = 0$
c) it is equal 1 for the sure event: $\Pr(\Omega) = 1$,
d) it complements its opposite event: $\Pr(A^C) = 1 - \Pr(A)$.

The above is known as **basic properties of the probability**. We will see later that these properties are satisfied by Kolmogorov probability (more preceisely: some of them stay in its definition).

Directly from the definition and above properties one can deduce that

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Hence, if events A and B are **disjoint** (i.e., they have no common outcomes, $A \cap B = \varnothing$) then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B).$$

The above extends for more than two events, e.g.

$$\begin{aligned}\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C) \\ - \Pr(A \cap B) - \Pr(B \cap C) - \Pr(C \cap A) \\ + \Pr(A \cap B \cap C),\end{aligned}$$

and

$$\Pr(A \cup B \cup C) = \Pr(A) + \Pr(B) + \Pr(C)$$

if A, B and C are **mutually disjoint**, i.e., $A \cap B = B \cap C = C \cap A = \varnothing$.

The above identities are known as an **addition rule for the probability**.

Example (***students at PUT***). Poznań University of Technology (PUT) is the 6th biggest higher engineering school in Poland. In 2012

PUT provided stationary studies to S =14.9 thousand students,
N = 5.7 thous. were educated in a non-stationary way,
P = 1.4 thous. participated in post-graduate studies
and D = 607 took doctoral studies.

Taking 600 instead of 607 we see that the total number of students was T = 22.6 thousand, the percentage share of these groups was

65.9, 25.2, 6.2 and 2.7%,

resp. Supposing that all these four groups are exclusive (i.e., there are no students who participate in different foms) we see that the probability a randomly chosen student is that of stationary, non-stationary, post-graduate and doctoral study is equal to

$$\Pr(S) = \frac{\mathrm{S}}{\mathrm{T}} = \frac{14.9}{22.6} = 0.659,\ \Pr(N) = \frac{\mathrm{N}}{\mathrm{T}} = \frac{5.7}{22.6} = 0.252,$$

$$\Pr(P) = \frac{\mathrm{P}}{\mathrm{T}} = \frac{1.4}{22.6} = 0.62,\ \Pr(D) = \frac{\mathrm{D}}{\mathrm{T}} = \frac{0.6}{22.6} = 0.027,$$

resp., where

S, N, P, D – (the event that) a randomly chosen student frequents stationary, non-stationary, post-graduate and doctoral study, resp.,

U – (the event that) a randomly chosen student frequents undergraduate study, so $U = S \cup N$.

We ask what is the probability $\Pr(U)$ of U. By the definition, it is the relation of two numbers: the number of events realizing U and the number of all possible cases; therefore

$$\Pr(U) = \frac{\mathrm{U}}{\mathrm{T}} = \frac{\mathrm{S}+\mathrm{N}}{\mathrm{T}} = \frac{14.9+5.7}{22.6} = \frac{20.6}{22.6} = 0.0912 = 91.2\%.$$

At the same time it is the sum od two probabilities, that of S and that of N, so

$$\Pr(U) = \Pr(S) + \Pr(N) = \frac{14.9}{22.6} + \frac{5.7}{22.6} = \frac{20.6}{22.6} = 91.2\ \%.$$

In fact, there are students taking education in at least two courses. Let's say that they are 400 persons, and 100 of them study, at the same semester, both in stationary and non-stationary way. Then the probability

$$\Pr(U) = \Pr(S) + \Pr(N) - \Pr(S \cap N) = \frac{14.9}{22.6} + \frac{5.7}{22.6} - \frac{0.1}{22.6} = \frac{20.5}{22.6} = 0.907 = 90.7\ \%.$$

Example ended.

Conditional probability

Given a finite set Ω and its nonempty subset A, the formula

$$\mathrm{Q}(B, A) := \frac{\Pr(A \cap B)}{\Pr(A)}$$

defines, for every subset B in A, the function Q. It is very easy to see that this is nothing else that the probability of the event B which is the subset in A. This probability is called a **conditional probability**, the probability of B conditioned by the occurrence of A. It is denoted by $\Pr(B|A)$, so

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)},$$

and read as 'the probability of the event B given that the event A has occurred',
'the probability of B given A'.

Example (***on conditional probability***). In a bag there are five red balls and three green balls. One ball is randomly drawn and set aside, next a second ball is drawn at random. We can denote the events:

A – the ball first drawn is red,
B – the ball drawn as the second one is red.

Then $A \cap B$ – both drawn balls are red,
$B|A$ – the second drawn ball is red provided the first selected ball is red (in other words the event is following 'from a beg where there are 4 red balls and 3 green balls there is drawn a red ball),

$$\Pr(A) = \frac{5}{8} = 0.625,\ \Pr(A \cap B) = \frac{\binom{5}{2}}{\binom{8}{2}} = \frac{5}{14} = 0.357,\ \Pr(B|A) = \frac{\binom{4}{1}}{\binom{7}{1}} = \frac{4}{7} = 0.571.$$

It verifies the formula for the conditional probability:

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\frac{5}{14}}{\frac{5}{8}} = \frac{8}{14} = \frac{4}{7}.$$

Example ended.

The formula for the conditional probability can be written as

$$\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B)$$

and it is called a **multiplication rule**.

We say that sets $B_1, B_2, \ldots, B_n$ cover a set A, or that $\{ B_1, B_2, \ldots, B_n \}$ is a **cover** of A, if $B_1 \cup B_2 \cup \ldots \cup B_n = A$. If these sets are mutually disjoint[6] then we say that the collection $\{ B_1, B_2, \ldots, B_n \}$ is a **proper cover**, or a (proper) **decomposition**, or a **partition**, of A.

Obviously, if $\{ B_1, B_2, \ldots, B_n \}$ is the partition of A, then

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)$$

and $$(A \cap B_j) \cap (A \cap B_k) = \varnothing \text{ for } j \neq k.$$

Therefore

$$\begin{aligned} \Pr(A) &= \Pr((A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n)) = \\ &= \Pr(A \cap B_1) + \Pr(A \cap B_2) + \ldots + \Pr(A \cap B_n) = \\ &= \Pr(A|B_1) \cdot \Pr(B_1) + \Pr(A|B_2) \cdot \Pr(B_2) + \ldots + \Pr(A|B_n) \cdot \Pr(B_n) \end{aligned}$$

(the last equality is validated by the multiplication rule).

This way we derived following **law of total probability**:

If $\{ B_1, B_2, \ldots, B_n \}$ is the partition of A, then

$$\Pr(A) = \Pr(A|B_1) \cdot \Pr(B_1) + \Pr(A|B_2) \cdot \Pr(B_2) + \ldots + \Pr(A|B_n) \cdot \Pr(B_n).$$

In particular case, when A is decomposed in two sets, $A = B \cup C$ and $B \cap C = \varnothing$, the law of total probability reads as the formula

$$\Pr(A) = \Pr(A|B) \cdot \Pr(B) + \Pr(A|C) \cdot \Pr(C).$$

This is the formula we use in the below example.

Example (***the second ace in the deck***). We want to know what is the probability the second card in the well shuffled card deck is an ace.
Let's denote: A_1 – the first card is an ace,
A_2 – the second card is an ace.
Then A_2^C means; 'the second card is not an ace'
and – by the formula for the total probability (with $A_1 = A_1 \cup A_1^C$) – we have

$$\begin{aligned} \Pr(A_2) &= \Pr(A_2|A_1) \cdot \Pr(A_1) + \Pr(A_2|A_1^C) \cdot \Pr(A_1^C) = \\ &= \frac{3}{51} \cdot \frac{4}{52} + \frac{4}{51} \cdot \frac{48}{52} = \frac{1}{13}. \end{aligned}$$

Example ended.

[6] we say that sets $B_1, B_2, \ldots, B_n$ are **mutually disjoint** if $B_j \cap B_k = \varnothing$ for $j \neq k$.

Bayes' formula

If $\{ B_1, B_2, \ldots, B_n \}$ is the partition of A, then, by the formula for the conditional probability applied twice, we have

$$\Pr(B_j|A) = \frac{\Pr(B_j \cap A)}{\Pr(A)} = \frac{\Pr(A \cap B_j)}{\Pr(A)} = \frac{\Pr(A \mid B_j) \cdot \Pr(B_j)}{\Pr(A)},$$

and, by the law of total probability (applied to the denominator),

$$\Pr(B_j|A) = \frac{\Pr(A \mid B_j) \cdot \Pr(B_j)}{\sum_{k=1}^{n} \Pr(A \mid B_k) \cdot \Pr(Bk)}.$$

This is called a **Bayes' formula**, or **Bayes' rule**, and the statement saying that this formula holds true is known as **Bayes' theorem** [7)].

Example (***a woman attending an conference***).
Among 100 persons attending an conference there are 50 physicists, 30 mathematicians and 20 computer specialists. Among them there are 20, 18 and 2 women, resp. We see a woman and, without asking her or nobody else, we would guess what is her profession. Let's denote

F, M, I = a randomly met person is a physicist, a mathematician, a IT specialist, resp.,

W = a randomly chosen participant of the conference is a woman.

In these denotations (of four random events) the data are:

$$\Pr(F) = \frac{50}{100} = 0.5,\ \Pr(M) = \frac{30}{100} = 0.3,\ \Pr(I) = \frac{20}{100} = 0.2,$$

$$\Pr(W|F) = \frac{20}{50} = 0.4,\ \Pr(W|M) = \frac{18}{30} = 0.6,\ \Pr(W|I) = \frac{2}{20} = 0.1.$$

Last three numbers are nothing else than the probabilities that a person blindly chosen from among physicists, among mathematicians and among IT specialists is a woman is

$$\Pr(W|F) = \frac{20}{50} = 0.4,\ \Pr(W|M) = \frac{18}{30} = 0.6,\ \Pr(W|I) = \frac{2}{20} = 0.1, \text{ resp.}$$

Directly form the data we also see that in the number of women participating in the meeting is 18 + 20 + 2 = 40,so the probability a blindly chosen participant is a woman is $\Pr(W) = 40/100 = 0.4$.

The probabilities that a woman is a physicist, a mathematician and an IT specialist is

---

[7)] The name recalls Thomas Bayes (1701-61), who first suggested using the theorem to update beliefs (that a considered event takes place). It was independently discovered by P.-S.Laplace who formulated it in his *Théorie analytique des probabilities* (1812).

$$\Pr(F|W) = \frac{\Pr(F \cap W)}{\Pr(W)} = \frac{20}{40} = 0.5\,,\ \Pr(M|W) = \frac{\Pr(M \cap W)}{\Pr(W)} = \frac{18}{40} = 0.45\,,$$

$$\Pr(I|W) = \frac{\Pr(I \cap W)}{\Pr(W)} = \frac{2}{40} = 0.05\,,\ \text{respectively}$$

Above calculated probabilities $\Pr(W)$, $\Pr(F|W)$, $\Pr(M|W)$ and $\Pr(I|W)$ are obtained, in fact, via the formula for the total probability and the Bayes' formula as follows

$$\Pr(W) = \Pr(W|F) \cdot \Pr(F) + \Pr(W|M) \cdot \Pr(M) + \Pr(W|I) \cdot \Pr(I) =$$
$$= \quad 0.6 \cdot 0.3 \quad + \quad 0.4 \cdot 0.5 \quad + \quad 0.1 \cdot 0.2 = 0.40,$$

$$\Pr(F|W) = \frac{\Pr(W \mid F) \cdot \Pr(F)}{\Pr(W)} = \frac{0.4 \cdot 0.5}{0.4} = 0.5\,,$$

$$\Pr(M|W) = \frac{\Pr(W \mid M) \cdot \Pr(M)}{\Pr(W)} = \frac{0.6 \cdot 0.3}{0.4} = 0.45\,,$$

$$\Pr(I|W) = \frac{\Pr(W \mid I) \cdot \Pr(I)}{\Pr(W)} = \frac{0.1 \cdot 0.2}{0.4} = 0.05\,.$$

The above values can be produced immediately by Bayes' formula applied with $n = 3$, $A = W$, $B_1 = F$, $B_2 = M$ and $B_3 = I$; really, we have:

$$\Pr(F|W) = \frac{\Pr(W \mid F) \cdot \Pr(F)}{\Pr(W \mid F) \cdot \Pr(F) + \Pr(W \mid M) \cdot \Pr(M) + \Pr(W \mid I) \cdot \Pr(I)} =$$
$$= \frac{0.4 \cdot 0.5}{0.4 \cdot 0.5 + 0.6 \cdot 0.3 + 0.1 \cdot 0.2} = \frac{0.2}{0.4} = 0.5\,,$$

$$\Pr(F|M) = \frac{\Pr(W \mid M) \cdot \Pr(M)}{\Pr(W \mid F) \cdot \Pr(F) + \Pr(W \mid M) \cdot \Pr(M) + \Pr(W \mid I) \cdot \Pr(I)} =$$
$$= \frac{0.6 \cdot 0.3}{0.4 \cdot 0.5 + 0.6 \cdot 0.3 + 0.1 \cdot 0.2} = \frac{0.18}{0.40} = 0.45\,,$$

$$\Pr(F|I) = \frac{\Pr(W \mid I) \cdot \Pr(I)}{\Pr(W \mid F) \cdot \Pr(F) + \Pr(W \mid M) \cdot \Pr(M) + \Pr(W \mid I) \cdot \Pr(I)} =$$
$$= \frac{0.1 \cdot 0.2}{0.4 \cdot 0.5 + 0.6 \cdot 0.3 + 0.1 \cdot 0.2} = \frac{0.02}{0.40} = 0.05\,.$$

Example ended.

In particular case (n=2: $A = B \cup B^{C}$) the Bayes' theorem says that

$$\Pr(B|A) = \frac{\Pr(A \mid B)}{\Pr(A)} \cdot \Pr(B)\,.$$

This formula is also derived directly from the multiplication rule; indeed, there is $\Pr(A \cap B) = \Pr(A|B) \cdot \Pr(B)$ and $\Pr(B \cap A) = \Pr(B|A) \cdot \Pr(A)$, so

$$\Pr(B|A) \cdot \Pr(A) = \Pr(A|B) \cdot \Pr(B)$$

and, for left sides are equal, it validates the above formula.

Bayesian updating

It is common to think of the Bayes' formula in terms of **updating the belief** in a hypothesis when new data arrive, new data are taken into account. To facilitate the memorization of this approach, let's write H and E instead of B and A, respectively. Then the Bayes' formula takes form

$$\Pr(H|E) = \frac{\Pr(E \mid H)}{\Pr(E)} \cdot \Pr(H)$$

and can be read in following terms:

- H – a hypothesis (aka an **initial belief**),
- E – an evidence (by definition, it is the collection of outcomes,
  and it is used to support or to reject the hypothesis H,
  to update the belief about a hypothesis H),
- $\Pr(H)$, called a **prior probability** that H holds true,
  is the probability of H before E is observed,
  is the initial degree of belief in A,
- $\Pr(E)$ – the probability E takes place
  (usually it is the frequency in which E occurred),
- $\Pr(H|E)$, called a **posterior probability** that H holds,
  is the probability of H after E is observed
  is the probability of H given E,
  is the degree of belief in A having accounted for B,
- $\Pr(E|H)$ – the likelihood that E occurs if H is true,
  the probability E occurs if H is true.

Using above terms we have a **Bayesian interpretation**, an **epistemological interpretation**:

the probability measures a degree of belief,

the Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence (this is: before the data are taken into account and after it)

it changes $\Pr(H)$ by the factor

$$\frac{\Pr(E \mid H)}{\Pr(E)}$$

representing the support the evidence E provides for the hypothesis H,
the impact of E on the probability of H;

this factor is sometimes called a **likelihood impactness** (of E on H).

This way, and in a concise form, the Bayes' formula can be read as follows:

posterior is linearly proportional to prior,
and the proportionality coefficient is the likelihood impactness.

Example (***cancer-ill rate***). We are interested in diagnosing cancer in patients who visit a chest clinic.
Let the hypothesis H = 'person has cancer'
the evidence E = 'person is a smoker'.
On the basis of past data we know
the prior probability $\Pr(H) = 0.1$
(10% of patients entering the clinic turn out to have cancer),
the evidence probability $\Pr(E) = 0.5$
(a half of patients smokes),
the posterior probability $\Pr(E|H) = 0.8$
(in every 10 suffering from cancer there are 8 smokers).
Therefore,
the likelihood impactness is $\frac{\Pr(E \mid H)}{\Pr(E)} = \frac{0.8}{0.5} = 1.6$
and, by Bayes' formula, the probability that a smoker will be diagnosed that suffers from cancer

$$\Pr(H|E) = \frac{\Pr(E \mid H)}{\Pr(E)} \cdot \Pr(H) = \frac{0.8}{0.5} \cdot 0.1 = 0.16.$$

As we see, the prior probability 0.1 (recall: it states that a person has cancer) is revised to a posterior probability of 0.16 (it says: 16% of smokers suffer from cancer). This is a significance increase: the probability a smoker is cancer-ill is 60% higher than an average cancer rate among all visitors.
Example ended.

In particular case, when A is decomposed in two sets, $A = B \cup C$ and $B \cap C = \varnothing$, the Bayes' formula is

$$\Pr(B|A) = \frac{\Pr(A \mid B) \cdot \Pr(B)}{\Pr(A \mid B) \cdot \Pr(B) + \Pr(A \mid C) \cdot \Pr(C)}.$$

This is the formula we use in the below example.

Example (***screening for a rare disease***, http://people.reed.edu/~jones/Courses/P02.pdf).
A blood test for a disease has the sensivity $\Pr(P|S) = 0.99$,
and the specificity $\Pr(P^C|S^C) = 0.99$,
where P and S are events: P = 'test results positively' =
'test shows that the patient has disease' =
'test finds that the patient is ill',
S = 'the patient has disease' = 'the patient is sick'.

Then $P^C = N$ = 'test does not detect the disease' =
'test results negatively' = 'test finds the patient is healthy',
$S^C = H$ = 'the patient is not sick' = 'the patient is healthy' =
'the patient has no disease' = 'the patient is disease-free',
$P|S^C = P|H$ = 'the test detects that the healthy patient is ill' =
'the healthy is recognized by the test for being sick',
$P^C|S = N|S$ = 'the test detects that the healthy patient is ill' =
'the sick is recognized by the test for being healthy'.

Obviously, the total population, T, can be decomposed in two ways:
$T = H \cup S$ – the sum of healthy people and sick persons,
$T = N \cup P$ – the sum of persons who are negatively tested and these who are positively tested.

There are given $\Pr(P|S) = 0.99$,
$\Pr(N|H) = 0.99$,
and we immediately calculate
$\Pr(P|H) = \Pr(P|S^C) = 1 - \Pr(P^C|S^C) = 1 - 0.99 = 0.01$.
$\Pr(N|S) = \Pr(P^C|S) = 1 - \Pr(P|S) = 1 - 0.99 = 0.01$.

Let's say that the prevalence [8] $\Pr(S) = 0.01$;
it is interpreted that among a 1000 people
about 990 persons are healthy and 10 persons are sick.

[8] In medicine, a **prevalence (proportion)**, or **prevalence rate**, (usually expressed as a percentage) is the proportion of individuals in a population found to have a disease (or other characteristic) under investigation. Prevalence is a statistical concept referring to the number of cases of a disease that are present in a particular population at a given time, whereas an **incidence** refers to the number of new cases that develop in a given period of time. For example, in 2009 in the USA about 600 thousand Americans died from the heart disease (that's 1 in very 4 death), so the prevalence of the death from the heart disease is 25% when related to the deaths, and 0.2% when related to the whole population (about 318 millions: 0.6/318=0.0019)

We ask what is the probability that positively tested patient is truly sick;
this is, we ask for $\Pr(S|P)$.

By Bayes' formula (used with $A = P$, $B = S$ and $C = H$) we have

$$\Pr(S|P) = \frac{\Pr(P \mid S)\cdot\Pr(S)}{\Pr(P \mid S)\cdot\Pr(S) + \Pr(P \mid H)\cdot\Pr(H)} = \frac{0.99\cdot 0.01}{0.99\cdot 0.01 + 0.01\cdot 0.99} = \frac{1}{2} = 50\%;$$

it says:

- the probability that a patient, who is recognized by the test for being sick, has the disease is ½;
- with the probability equal to ½ the positively tested person suffers from the disease,
- roughly a half of those who test positively has the disease.

In the same way we obtain that

$$\Pr(S|P) = \frac{\Pr(P \mid S)\cdot\Pr(S)}{\Pr(P \mid S)\cdot\Pr(S) + \Pr(P \mid H)\cdot\Pr(H)} = \frac{0.99\cdot 0.01}{0.99\cdot 0.01 + 0.01\cdot 0.99} = \frac{1}{2} = 50\%;$$

It implies immediately that $\Pr(H|P) = 1 - \Pr(S|P) = 0.5$;
it says:

- the probability that a patient, who is recognized by the test for being sick, has no disease is ½;
- with the probability equal to ½ the positively tested person does not suffers the disease.

So, in our numerical example, where 1000 people are taken into account, among positively tested persons
the disease is detected in 50% of ill people,
and in 50% of healthy people;
in absolute numbers:
among 10 ill persons the test detects the disease in 5 persons,
among 990 healthy persons the test detects that 495 persons are disease-free.

Example ended.

Below we give three examples which concern the quality of the control. Every control

- tests some objects (an article produced in a factory and examined for meeting the quality standards, a persons tested to recognize a disease) underwent some risk,
- is realized with some sensitivity and specificity,
- intentionally is done to detect bad items (such as articles of poor quality, persons suffering from an illness).

As it was already mentioned, the level of the risk (that a product is poor, is out the required quality, that a person is sick, etc.) is called a prevalence. The prevalence can be stated by data recorded in previous years (the defectiveness in previous years, a number of people suffering from the disease in last years, etc.) and is believed it remains the same at the current moment. In many cases the prevalence is calculated from the current situation.

Let's deal on with the people. Not to bother about the sex (it requires to operate different pronouns: he, she, his, her) let's say that we deal with men only.

A man is either sick or healthy; the states 'sick' and 'healthy' are his true states (sometimes we say: 'truly sick', 'truly healthy'), we will denote them by letters *S* and *H*, resp.

A test provides either positive or negative result; we will denote these results by letters *P* and *N*, resp.

There are possible four outcomes:

true positive: *S* and *P* (a man is sick, the test shows it),
false positive: *H* and *P* (a man is healthy, the test shows he sick),
false negative: *S* and *N* (a man is sick, the test does not detect his illness),
true negative: *H* and *N* (a man is healthy, the test does not diagnose it, the test shows he is sick).

This all can be presented in the table (let's call it a **pair-matching table**) as follows:

	the true state		
	sick (*S*)	healthy (*H*)	
positive test (*P*)	$S \cap P$	$H \cap P$	
negative test (*N*)	$S \cap N$	$H \cap N$	

Let *sp*, *hp*, *sn* and *hn* stand for numerical values staying in respective fields of this table (e.g., $st = sp + sn$ is the total number of sick persons who are detected either as sick or as healthy; *hn* is the number of persons who are healthy and are negatively tested). These four denotations, and five more ones (*st*, *ht*, *tp*, *nt* and *t*) are present in the table (called a **2×2 contingency table**, a **cross tabulation**, a **cross tab matrix**, a **cross frequency matrix**) below.

	the true state		
	sick (*S*)	healthy (*H*)	
positive test (*P*)	*sp*	*hp*	$pt = sp + hp$
negative test (*N*)	*sn*	*hn*	$nt = sn + hn$
	$st = sp + sn$	$ht = hp + hn$	*t*

The five new quantities (*st*, *ht*, *tp*, *nt* and *t*) are defined as follows:

the total number of sick persons: $st = sp + sn$,
the total number of healthy persons: $ht = hp + hn$,
the total number of persons positively tested: $pt = sp + hp$,
the total number of persons negatively tested: $nt = sn + hn$,
the total number of persons: $t = st + ht$
$= pt + nt$
$= sp + sn + hp + hn$.

There can be calculated some new quantities. We start with

the **sensitivity of the test**: $\text{Sens} := \frac{sp}{st} = \frac{sp}{sp + sn}$,

the **specificity of the test**: $\text{Spec} := \frac{hn}{ht} = \frac{hn}{hp + hn}$.

The pair of these quantities, the sensitivity and the specificity, is said to be **basic characteristics of the test**. Usually, they are expressed in percents.

As it was told above, in many cases the prevalence is calculated from the current data; this means the prevalence is

$$\text{Prev} := \frac{st}{t} = \frac{st}{st + ht} = \frac{st}{sp + sn + hp + hn}.$$

The sensitivity of the test, Sens, is the fraction of sick persons who are positively tested, it is interpreted as

the probability the test indicates a disease among sick persons,
the probability the test signals that an ill man is ill,
the probability of a positive test, given that a man is sick.

In other words, the sensitivity of the test, Spec, is the ability of this test to correctly identify persons suffering from the disease.

The specificity of the test, Spec, is the fraction of healthy persons who are negatively tested, it is interpreted as

the probability that the test recognizes a healthy person as disease-free,
the probability the test diagnoses correctly the good condition of a man,
the probability of a negative test, given that a man is well.

The specitivity of the test is the ability of this test to correctly identify persons who are disease-free.

Both sensitivity and specificity are quantities which do not depend on the prevalence of the sickness in a population. They both are used to have

false positive rate, or a **type I error**: $\alpha := 1 - \text{Spec} = \frac{hp}{ht} = \frac{hp}{hp + hn}$,

false negative rate, or a **type II error**,

or a **power** of the test: $\beta := 1 - \text{Sens} = \frac{sn}{st} = \frac{sn}{sp + sn}$,

(positive) **likehood ratio**: $\frac{\beta}{\alpha} = \frac{\text{Sens}}{1 - \text{Spec}}$,

a (**rand**) **accuracy**: $\text{Prev} \cdot \text{Sens} + (1 - \text{Prev}) \cdot \text{Spec} = \frac{sp + hn}{t}$.

Besides two characteristics of the test there can also be calculated

the **positive predicted value** (**ppv**): $\frac{sp}{pt} = \frac{sp}{sp + hp}$

the **negative predicted value** (**npv**): $\frac{hn}{nt} = \frac{hn}{sn + hn}$.

In the information retrieval the ppv is called a **precision**,
and the sensitivity is called a **recall**.

Above quantities can be arranged in the table, called a 2×2 **confusion matrix**, or an **error matrix**, which differs from that considered above only by listing not absolute values but relative quantities (usually expressed in percent). The crucial elements of the confusion matrix are the errors of the 1st type (α) and the error of the 2nd type (β), two other elements are their complements ($\text{Sens} = 1 - \alpha$, $\text{Spec} = 1 - \beta$), and the confusion matrix has the form:

		the true state:		
		sick (*S*)	healthy (*H*)	
test outcome:	positive (*P*)	$\text{Sens} = 1 - \alpha$	α	
	negative (*N*)	β	$\text{Spec} = 1 - \beta$	

All four values (the sensitivity, the specificity, ppv and npv) are used to measure correctness of the test.

We have a **good test** if its sensitivity and its specificity are high. Indeed,

the sensitivity is related to the detection of the disease
(an ill man is diagnosed as ill),

the specificity is related to the detection of the healthy condition
(a disease-free man is tested negatively,
a healthy man is diagnosed as healthy).

Example (***530 person tested for an illness***).
530 people are tested for an illness. 53 people suffer from this sickness. So 530–53=477 people do not suffer from this diseased, they are not affected by this illness, they stay healthy, disease-free, and the prevalence is

$$\text{Prev} = \frac{53}{530} = 10\%.$$

Among these 530 persons there are

sp = 50 diseased and positively tested,
sh = 10 diseased, but not diagnosed as sick,
sn = 3 disease-free and diagnosed as sick,
hn = 477 diseased-free and negatively tested.

These data fill the below cross frequency table; moreover, there are also placed the quantities *st*, *ht* (calculated by summing in columns), *tp*, *nt* (calculated by summing in rows) and *t* (the sum of all persons).

	the true state		
	sick (*S*)	healthy (*H*)	
positive test (*P*)	50	10	60
negative test (*N*)	3	467	470
	53	477	525

So the sensitivity of the test is $\text{Sens} = \frac{sp}{st} = \frac{50}{53} = 94.3\%$

(it says: the test correctly diagnoses the sick in 943 cases per each 1000; in 94.3% the person diagnosed as ill is really ill),

the specitivity of the test is $\text{Spec} = \frac{hn}{ht} = \frac{467}{477} = 97.9\%$

(it says: the test does not recognize the illness in 979 cases in every 1000 healthy patients; in 97.9% the person diagnosed as healthy is disease-free),

ppv (the positive predicted value) is $\frac{sp}{tp} = \frac{50}{60} = 8.3\%$

(when the test diagnoses the illness, the reliability of this diagnosis is 8.3%),

npp (the negative predicted value) is $\frac{hn}{nt} = \frac{467}{470} = 99.3\%$

(we believe that 99.3% of persons with the disease not detected is disease-free).

Note that the corresponding error matrix is

		the true state:		
		sick (*S*)	healthy (*H*)	
test outcome:	positive (*P*)	Sens = 94.3	α = 2.1	
	negative (*N*)	β = 5.7	Spec = 97.9	

It's clear that all above can be presented not only in the health-ill terms (which are related to people), but also in the good-bad, or defective-correct terminology (which is related to the quality of the products made in a factory), in the positive-negative duality (which is used to describe economical options) etc. in the example below we deal with the quality control, in particular we will recognize the sensitivity of the control (device). It is usually expressed in percents, it is the correctness of the recognition. We will find that

the control sensitivity is 99.7%; it means that 99.7% of good products are recognized (or: classed) by the control as good, that in 1000 articles classed as good ones there are, at average, 997 correct articles (and 3 articles which do not satisfy the good quality requirements),

the control specificity is 35.8%; it means that of defect products (90% of defect products are recognized correctly, i.e., the quality control detects 90% of bad products as defect ones).

Example (***the quality of the quality control***).
The sensivity and the specificity of QuaCo (= the quality control) are 95% and 90%, resp.; this means that the correctness of the recognition is 95% of good products (95% of good products are recognized as good ones) and 90% of defect products (90% of defect products are recognized correctly, i.e., the quality control detects 90% of bad products as defect ones). Previous controls showed that the prevalence of defect products is 3%; this means that it is assumed that 3% of total production is defective.
Let's designate the events:

G – a product chosen at random is of good quality,

D – a randomly selected product is defective,

A – a randomly controlled product appears to be of good quality, QuaCo approves it as a good one,

R – QuaCo classes a product for defective (and, after the rules stated in the factory, this product is rejected from sending to the market).

In these symbols
the data say:

$\Pr(D) = 0.03$ (it is assumed that 3% of elements are of poor quality),

$\Pr(A|G) = 0.95$ (the sensivity of QuaCo: QuaCo classes 95% of examined good products for good),

$\Pr(R|D) = 0.90$ (the specivity of QuaCo: QuaCo classes 90% of examined bad products for bad),

so $\Pr(G) = 1 - \Pr(D) = 0.97$ (97% of elements are of good quality),

$\Pr(R|G) = 1 - \Pr(A|G) = 0.05$ (5% of good elements are classed as wrong),

$\Pr(A|D) = 1 - \Pr(R|D) = 0.10$ (10% of poor elements are classed as good)

and

$\Pr(G|A)$ is the probability that a product classes by QuaCo for good is good,

$\Pr(D|R)$ is the probability that a product classes by QuaCo for poor is poor.

The last two probabilities are that we want to know. They can be calculated via Bayes' formula as follows

$$\Pr(G|A) = \frac{\Pr(G \cap A)}{\Pr(A)} = \frac{\Pr(A \cap G)}{\Pr(A)} = \frac{\Pr(A \mid G) \cdot \Pr(G)}{\Pr(A \mid G) \cdot \Pr(G) + \Pr(A \mid D) \cdot \Pr(D)} =$$

$$= \frac{0.95 \cdot 0.97}{0.95 \cdot 0.97 + 0.10 \cdot 0.03} = \frac{0.9215}{0.9245} = 0.997,$$

$$\Pr(D|R) = \frac{\Pr(D \cap R)}{\Pr(R)} = \frac{\Pr(R \cap D)}{\Pr(R)} = \frac{\Pr(R \mid D) \cdot \Pr(D)}{\Pr(R \mid D) \cdot \Pr(D) + \Pr(R \mid G) \cdot \Pr(G)} =$$

$$= \frac{0.90 \cdot 0.03}{0.90 \cdot 0.03 + 0.05 \cdot 0.97} = \frac{0.027}{0.0755} = 0.358.$$

These results say:

- the probability that a product attested, by QuaCo, as of good quality is so is equal to 99.7%; in 99.7 percent a product recognized by QuaCo as good is good,
- the probability that a product recognized by QuaCo as defective is of low quality is equal to 35.8%; QuadCo classes a poor product as rejective with the certainty equal to 35.8%.

Hence we immediately get

$\Pr(D|A) = 1 - \Pr(G|A) = 1 - 0.997 = 0.003$
(QuaCo approves a defective product with the probability 0.3%),

$\Pr(G|R) = 1 - \Pr(D|R) = 1 - 0.358 = 0.642$
(the probability that an article recognized by QuadCo as poor is not so is equal to 64.2%).

Shortly saying, QuaCo is

- highly restrictive wrt defective products (it wrongly classes as little as 3 in 1000 poor articles; in other words: after the QuaCo in every 1000 articles approved to be sent to market there are 3 out of the quality, there are 3 ones which should be not sent to market),
- much more tolerant to good articles (in every 1000 articles classed as of poor condition there are 358 which are truly defective, and 642 which are of good quality).

see https://onlinecourses.science.psu.edu/stat507/node/71

Example (***medical test***).
Let's denote S – a person is sick,
H – a person is healthy,
P – a medical test results positively, i.e., it detects the person underwent the test has the disease,
N – a medical test results negatively, i.e., it does not recognize a person as sick,
and let's work with $\Pr(S) = 0.01$;
the test sensivity $\Pr(P|S) = 0.99$,
the test specificity $\Pr(P^C|S^C) = 0.99$,

it is interpreted that among a 1000 people about 990 persons are healthy and 10 persons are sick.
where P and S are events: P = 'test results positively' = 'test shows that the patient has disease' = 'test finds that the patient is ill',
S = 'the patient has disease' = 'the patient is sick'.
Then $P^C = N$ = 'test does not detect the disease' = 'test results negatively' = 'test finds the patient is healthy',
$S^C = H$ = 'the patient is not sick' = 'the patient has no disease', 'the patient is healthy',
$P|S^C = P|H$ = 'the test detects that the healthy patient is ill' = 'the healthy is recognized by the test for being sick',
$P^C|S = N|S$ = 'the test detects that the healthy patient is ill' = 'the sick is recognized by the test for being healthy'.
Obviously, the total population, T, can be decomposed in two ways:
$T = H \cup S$ – the sum of healthy people and sick persons,
$T = N \cup P$ – the sum of persons who are negatively tested and these who are positively tested.
There are given
$\Pr(P|S) = 0.99$,
$\Pr(N|H) = 0.99$,
and we immediately calculate
$\Pr(P|H) = \Pr(P|S^C) = 1 - \Pr(P^C|S^C) = 1 - 0.99 = 0.01$.
$\Pr(N|S) = \Pr(P^C|S) = 1 - \Pr(P|S) = 1 - 0.99 = 0.01$.

Let's say that the
We ask what is the probability that positively tested patient is really sick;
this is, we ask for $\Pr(S|P)$.

Example (***studying in Poland in 2011***). In 2011 the population of Poland was 38.5 mln, of which 52% were female. The tertiary education was provided to 1.77 mln, of which 55% were female. We will calculate is the probability that a citizien of Poland is a either female or student.

Realizing approximate arithmetics we see that in 2011 in Poland there were

$0.52 \cdot 38.5 = 20.0$ mln women,

$0.55 \cdot 1.77 = 0.97$ mln women undergoing the tertiary education.

The data (numbers 38.5, 20.0, 1.77 and 0.97) are put, in bold, in the below table. This table consists also 5 other numbers (36.73, 19.03, 18.5, 17.07 and 0.80) calculated (by substracting appropriate values) from the data.

in mln	on studies	other	total
female	**0.97**	19.03	**20.0**
male	0.80	17.70	18.5
total	**1.77**	36.73	**38.5**

Both groups, female and male, as well as person taking education at heis (hei = higher education institutions) and that do not, are exclusive, so, for instance,

$$\Pr(\text{a Pole was in 2011 educated at hei}) = \frac{1.77}{38.5} = 0.046 = 4.6\ \%,$$

and it is equal to the sum

Pr(a female Pole was in 2011 educated at hei)

+ Pr(a male Pole was in 2011 educated at hei)

$$= 0.097 \cdot 20.0 + 0.080 \cdot 18.5 = \frac{0.55 \cdot 1.77}{38.5}$$

When dealing with percentages, in 2011 in Poland

$0.97/20.0 = 4.85\%$ of female population underwent the tertiary education,

$0.80/18.5 = 4.32\%$ of male population took the studies

$1.77/38.5 = 4.60\ \%$ of citizens studied at heis,

$36.73/38.5 = 95.40\ \%$ of citizens studied at higher education institutions

(when calculating directly from percentage data we obtain that the tertiary education is provided to $(4.85+4.32)/2 = 4.585$ percentage of population, as is not to $(95.15+95.68)/2 = 95.415 = 100 - 4.585$ percentage).

in %	on studies	other	total
female	**4.85**	95.15	**52**
male	4.32	95.68	48
total	**4.60**	95.40	**100**