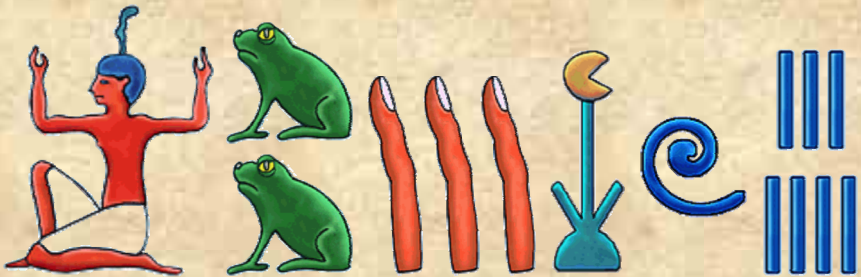




Boris Obsieger

NUMERICAL METHODS I

Basis and Fundamentals



$\overline{M} \overline{C} \overline{C} \overline{X} \overline{X} \overline{X} M C V I I$

1 0010 1100 1001 0000 0011

Including 76 examples and 13 algorithms













ISBN 978-953-7919-25-2
In **colour** by Ingram Digital

1. NUMERAL SYSTEMS

There are three basic concepts of a number representation:

- sign-value notation,
- subtractive sign-value notation,
- positional notation (place-value notation).

A *sign-value notation* represents numbers by a series of numerals whose sum show up the represented number. In the Ancient Egyptian numeral system, for example,  means hundred and  means ten (Tab. 1.1), so  means three hundred and ten ($100 + 100 + 100 + 10$).

Tab. 1.1. Ancient Egyptian numerals				
Value	1	10	100	1000
Hieroglyph				
Description	Single stroke	Heel bone	Coil of rope	Water lily (Lotus)
Value	10 000	100 000		1 000 000
Hieroglyph				
Description	Finger	Tadpole or Frog		Man with both hands raised

27 hieroglyphs (𐀀𐀁𐀂𐀃𐀄𐀅𐀆𐀇𐀈𐀉𐀊𐀋𐀌𐀍𐀎𐀏𐀐𐀑𐀒𐀓𐀔𐀕𐀖𐀗𐀘𐀙𐀚𐀛).

Acrophonic Greek numerals

A little bit shorter form of sign-value notation was based on Acrophonic Greek numerals¹, innovated in the first millennium BC. Beside the symbols for 1, 10, 100, 1000 and 10000, the system had intermediate symbol for 5 and compound symbols for 50, 500, 5000 and 50000. The compound symbols were made by combining the symbol 5 with the symbols 10, 100, 1000 and 10000, Tab. 1.2.

Tab. 1.2. Acrophonic Greek numerals

Value	1	5	10	5·10	100	5·100	1000
Symbol	I	Γ or Π	Δ	ℱ	H	ℱ ^h	X
Value		5·1000	10 000	5·10 000	Example: 2064		
Symbol		ℱ	M	ℱ ^m	XXℱ ^Δ ΔIIII		

Alphabetic Greek numerals

Further improvements in a sign-value notation were achieved in the fifth century BC in the Alphabetic Greek numerals. Instead of using the separate set of symbols for numbers, values were assigned to lowercase letters of the old Greek alphabet based on their native alphabetic order, Tab. 1.3.

Tab. 1.3. Values assigned to symbols in the old Greek alphabet

Value	1	2	3	4	5	6	7	8	9
Symbol	α	β	γ	δ	ε	ς	ζ	η	θ
Value	10	20	30	40	50	60	70	80	90
Symbol	ι	κ	λ	μ	ν	ξ	ο	π	ρ
Value	100	200	300	400	500	600	700	800	900
Symbol	ρ	ς	τ	υ	φ	χ	ψ	ω	ξ
Value	1000	2000	3000	4000	5000	Example: 2064			
Symbol	,α	,β	,γ	,δ	,ε	,βξδ'			

¹ The reason why this system is called *acrophonic* is because the numerals for 5, 10, 100, 1000 are the first letters of the Greek words for these numbers, namely *ΠΕΝΤΕ*, *ΔΕΚΑ*, *ΗΕΚΑΤΟΝ*, and *ΧΙΛΙΑΙΟΝ*.

Of the 27 letters, nine were for units (1, 2, ..., 9), nine for tens (10, 20, ..., 90) and nine for hundreds (100, 200, ..., 900). To mark that a sequence of letters is in fact a number (and not a text), a special sign like a vertical dash or accent assign is placed after the numerals, i.e., σνζ´ which represents 200+50+7=257. To write numbers larger than 1000, a similarly vertical dash sign is placed before the numerals and below the line of writing, such as ,αλπθ´ representing 1000+900+80+9=1989.

Glagolitic numerals

The Glagolitic¹ numeral system is similar to the Alphabetic Greek numeral system. To Glagolitic letters were assigned values based on their native alphabetic order, Tab. 1.4. Nine letters were for units (1, 2, ..., 9), nine for tens (10, 20, ..., 90), nine for hundreds (100, 200, ..., 900) and remainder for thousands (1000, 2000,...). Numbers were distinguished from text by small square marks ♦, one before and one after each symbol.

Tab. 1.4. Values assigned to symbols in the Glagolitic alphabet ²									
Value	1	2	3	4	5	6	7	8	9
Symbol	ⱁ	ⱃ	ⱅ	ⱇ	ⱉ	ⱋ	ⱍ	ⱏ	ⱑ
Value	10	20	30	40	50	60	70	80	90
Symbol	ⱓ	ⱕ	ⱗ	ⱙ	ⱛ	ⱝ	ⱟ	Ⱡ	ⱡ
Value	100	200	300	400	500	600	700	800	900
Symbol	ⱄ	ⱆ	ⱈ	ⱊ	ⱌ	ⱎ	ⱐ	ⱒ	ⱔ
Value	1000	2000	3000	4000	5000	Example: 2064			
Symbol	ⱖ	ⱘ	ⱚ	ⱜ	ⱞ	♦ⱘ♦ⱎ♦ⱇ♦			

¹ The Glagolitic is the oldest known Slavic alphabet. It was invented during the 9th century by the Byzantine missionaries St Cyril (827-869 AD) and St Methodius (826-885 AD) in order to translate the Bible and other religious works into the language of the Great Moravia region. It is probably modelled on a cursive form of the Greek alphabet while their translations are based on a Slavic dialect of the Thessalonika area, which formed the basis of the literary standard known as Old Church Slavonic. This old Slavic scripts had remained in use by Croats up to 19th century.

² The Croation version of symbols which was used about 14th century .

For example, **h** denotes a letter, while **♦h♦** denotes the numeral “1”. If needed, to represent numbers which are not assigned to any letter, two or more symbols would have to be adjoined together. For example, the number 2014 may be written as $2000+4+10 = \text{♦M♦♦D♦♦}$.

Subtractive sign-value notation

Some improvements were made by *subtractive sign-value notation*. Roman numerals, for example, are generally written in descending order from left to right, but where a symbol of a smaller value precedes a symbol of a larger value, a smaller value is subtracted from a larger value, and the result is added to the total¹. For example, M means thousand and I means one (Tab. 1.5), so MI denotes number $1000+1=1001$, while IM denotes number $-1+1000=999$.

Tab. 1.5. Roman numerals ²							
Value	1	5	10	50	100	500	1000
Symbol	I	V	X	L	C	D	M
Value		5 000	10 000	50 000	100 000	500 000	1 000 000
Symbol		\overline{V}	\overline{X}	\overline{L}	\overline{C}	\overline{D}	\overline{M}

Note that there is no need for the zero in a sign-value notation.

Sign-value notation was the pre-historic way of writing numbers and only gradually evolved into the *positional notation*, also known as *place-value notation*, in which the value of a particular digit depends both on the digit itself and its position within the number.

Positional notation

In the *positional notation*, a number is represented by a sign (plus or minus) and digits, while the digital coma or the digital point (depending on a country) separates an integer part of a number from its fraction. Unlike the sign-value

¹ The notation of Roman numerals has varied through the centuries. Originally, it was common to use IIII instead of IV to represent four, because IV are the first two letters and consequently abbreviation for IVPITER, the Latin script spelling for the Roman god Jupiter. The notation which uses IV instead of IIII has become the standard notation in modern time, but with some exceptions. For example, Louis XIV, the king of France, who preferred IIII over IV, ordered his clockmakers to produce clocks with IIII and not IV, and thus it has remained [1].

² Roman numerals have remained in use mostly for the notation of *Anno Domini* years, and for numbers on clockfaces. Sometimes, Roman numerals are still used for enumerating the lists (as an alternative to alphabetical enumeration), and for numbering pages in prefatory material in books.

notation, there is an explicit need for zero in the positional notation. The number of digits with integer values zero, one, two, ... that form a positional numeral system is called the **base of the numeral system** (e.g. if there are 10 digits with values from zero to nine, the base is 10).

An integer number can be represented in a positional numeral system in base B with a sum of digits $d_k \in \{\text{zero, one, ..., } B-1\}$ multiplied by powers k of the base B as

$$inumber = (d_n \dots d_1 d_0)_B = d_n \cdot B^n + \dots + d_1 \cdot B^1 + d_0 \cdot B^0. \quad (1.1)$$

A real number can be represented in base B as

$$rnumber = (d_n \dots d_0, d_{-1} \dots)_B = d_n \cdot B^n + \dots + d_0 \cdot B^0 + d_{-1} \cdot B^{-1} + \dots. \quad (1.2)$$

In both cases the leading zeroes are usually omitted assuming that d_n is first non-zero digit.

Sexsagesimal system

Possibly the oldest positional numeral system is a **sexsagesimal system**, used around 3100 B.C., in Babylon. It is a combination of the sign-value and the positional notation in base 60 [2].

In Babylon, digits up to 59 were noted by using two symbols in the sign-value notation: symbol ∇ to count units and symbol \llcorner to count tens. These symbols ∇ and \llcorner and their values were combined to form 59 digits in a notation similar to that of Roman numerals; for example, the combination $\llcorner\llcorner$ represented the digit with a value of 23 (Fig. 1.1).

∇ 1	$\llcorner \nabla$ 11	$\llcorner \llcorner \nabla$ 21	$\llcorner \llcorner \llcorner \nabla$ 31	$\llcorner \llcorner \llcorner \llcorner \nabla$ 41	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla$ 51
$\nabla \nabla$ 2	$\llcorner \nabla \nabla$ 12	$\llcorner \llcorner \nabla \nabla$ 22	$\llcorner \llcorner \llcorner \nabla \nabla$ 32	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla$ 42	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla$ 52
$\nabla \nabla \nabla$ 3	$\llcorner \nabla \nabla \nabla$ 13	$\llcorner \llcorner \nabla \nabla \nabla$ 23	$\llcorner \llcorner \llcorner \nabla \nabla \nabla$ 33	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla$ 43	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla$ 53
$\nabla \nabla \nabla \nabla$ 4	$\llcorner \nabla \nabla \nabla \nabla$ 14	$\llcorner \llcorner \nabla \nabla \nabla \nabla$ 24	$\llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla$ 34	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla$ 44	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla$ 54
$\nabla \nabla \nabla \nabla \nabla$ 5	$\llcorner \nabla \nabla \nabla \nabla \nabla$ 15	$\llcorner \llcorner \nabla \nabla \nabla \nabla \nabla$ 25	$\llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla$ 35	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla$ 45	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla$ 55
$\nabla \nabla \nabla \nabla \nabla \nabla$ 6	$\llcorner \nabla \nabla \nabla \nabla \nabla \nabla$ 16	$\llcorner \llcorner \nabla \nabla \nabla \nabla \nabla \nabla$ 26	$\llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla \nabla$ 36	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla \nabla$ 46	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla \nabla \nabla \nabla$ 56
$\llcorner \nabla$ 7	$\llcorner \llcorner \nabla$ 17	$\llcorner \llcorner \llcorner \nabla$ 27	$\llcorner \llcorner \llcorner \llcorner \nabla$ 37	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla$ 47	$\llcorner \llcorner \llcorner \llcorner \llcorner \llcorner \nabla$ 57
$\llcorner \nabla \nabla$ 8	$\llcorner \llcorner \nabla \nabla$ 18	$\llcorner \llcorner \llcorner \nabla \nabla$ 28	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla$ 38	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla$ 48	$\llcorner \llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla$ 58
$\llcorner \nabla \nabla \nabla$ 9	$\llcorner \llcorner \nabla \nabla \nabla$ 19	$\llcorner \llcorner \llcorner \nabla \nabla \nabla$ 29	$\llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla$ 39	$\llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla$ 49	$\llcorner \llcorner \llcorner \llcorner \llcorner \llcorner \nabla \nabla \nabla$ 59
$\llcorner \llcorner$ 10	$\llcorner \llcorner \llcorner$ 20	$\llcorner \llcorner \llcorner \llcorner$ 30	$\llcorner \llcorner \llcorner \llcorner \llcorner$ 40	$\llcorner \llcorner \llcorner \llcorner \llcorner \llcorner$ 50	

Fig. 1.1. Digits used in Babylonian numeral system

These 59 digits are then used in positional numeral system in base 60 as it is illustrated in Example 1.1.

Example 1.1. Babylonian numerals:

$$\begin{aligned}
 \lllll &= 10 + 10 + 1 + 1 + 1 = 23, \\
 \lll &= 10 + 10 + 10 = 30, \\
 \ll &= 1 + 1 = 2, \\
 \llll \ll &= ("22" "1")_{60} = 22 \cdot 60^1 + 1 \cdot 60^0, \\
 \ll \lll &= ("2" "30")_{60} = 2 \cdot 60^1 + 30 \cdot 60^0.
 \end{aligned} \tag{1.3}$$

The Babylonians did not have a digit for zero. What the Babylonians used instead was a space (and later a disambiguating placeholder symbol \ll) to indicate a place without value, similar to zero.

In addition, Babylonians did not have any mark to separate integer from the fractional part of a number, but they calculated with real numbers, as it is illustrated in Example 1.2.

Example 1.2. The Babylonian approximation of the square root of 2 in the context of Pythagoras' theorem for an isosceles triangle.

The approximation is illustrated on the round tablet that was, we believe, an education tablet from Ancient Babylonia, dated 1800 B.C. [3]. The tablet, Fig. 1.2, has a square with both diagonals drawn in. On one side of the square is written $\lll = 30$, the length of the square side. If this is treated as a fraction of the number (i.e., as the first figure of the fraction), then

$$\lll = 30 \cdot 60^{-1} = 1/2. \tag{1.4}$$

Along one of the diagonals, the number is written

$$\ll \llll \lllll \ll = 1 + 24/60 + 51/60^2 + 10/60^3 = 1,41421296... \approx \sqrt{2} \tag{1.5}$$

and below it is the number

$$\lllll \llll \lllll = 42/60 + 25/60^2 + 35/60^3 = 0,70710648... \approx \sqrt{1/2}. \tag{1.6}$$

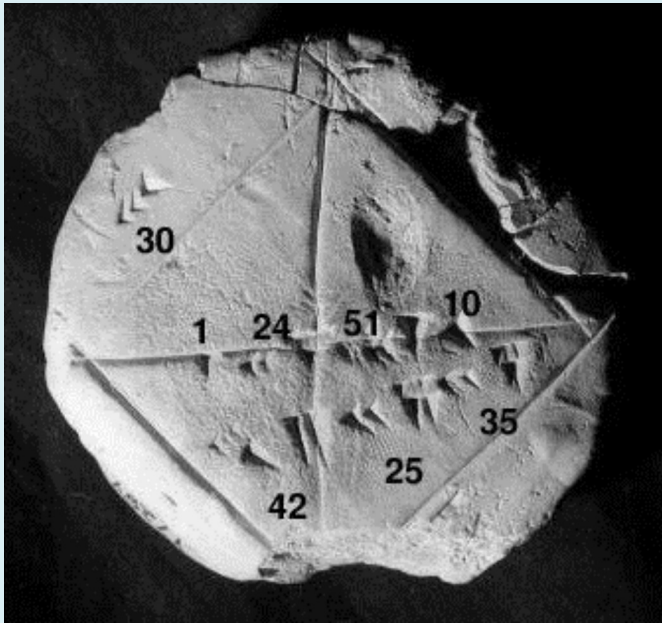


Fig. 1.2. Round tablet illustrating approximation of square root [3]











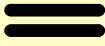



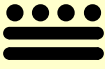




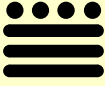
In fact, the number $\nabla \llcorner \llcorner \llcorner \llcorner \llcorner$ is a remarkably good approximation of $\sqrt{2} = 1,41421356\dots$ in four sexagesimal figures, which is about six decimal figures. Moreover, it is easy to see that $\nabla \llcorner \llcorner \llcorner \llcorner \llcorner \approx \sqrt{2}$ multiplied by $1/2$ is $\llcorner \llcorner \llcorner \llcorner \llcorner \approx \sqrt{1/2}$, the diagonal length of the square of the side $\llcorner = 1/2$.

Hexagesimal system




This system must be distinguished from the sexsagesimal system, although both systems have the same base (and that is 60). One digit in the hexagesimal system is a decimal number from 0 to 59. Today, the hexagesimal system is used to express time (hours, minutes and seconds).

Vigesimal system

Another interesting positional system is a *vigesimal system* (base-twenty) used by the Pre-Columbian Maya civilization. The Maya numerals were made up by combining three symbols in signed value notation: zero (shell shape \ominus), one (a dot \bullet) and five (a bar —). The construction of 20 numerals (starting from zero) is shown in Tab. 1.6. For example, seventeen (III) is written as two dots in a horizontal row above three horizontal bars stacked upon each other.

Tab. 1.6. Digits used in numeral system of Pre-Columbian Maya					
Value	0	1	2	3	4
Glyph					
Value	5	6	7	8	9
Glyph					
Value	10	11	12	13	14
Glyph					
Value	15	16	17	18	19
Glyph					

Because the base of the numeral system was 20, larger numbers were written down in powers of 20 from bottom to top. Fig. 1.3 shows how the number 2402 was written.



$6 \cdot 20^2 = 2400$
 $0 \cdot 20^1 = 0$
 $2 \cdot 20^0 = 2$
sum = 2402

Fig. 1.3. Number 2402 in Maya numeral system

As it can be seen, the addition is just a matter of adding up dots and bars. Maya merchants often used cocoa beans, which they laid out on the ground to do these calculations.

Addition is performed by combining the numeric symbols at each level

 $(7 + 6 = 13).$

(1.7)

If five or more dots result from the combination, then five dots are replaced by a bar. If four or more bars result from the combination, then four bars are replaced by a shell and a dot is added to the next higher row.

In subtraction, elements of the subtrahend symbol are removed from the minuend symbol:

$$\begin{array}{c} \text{---} \\ \text{☵} \end{array} \begin{array}{c} \text{---} \\ \text{☰} \end{array} = \begin{array}{c} \text{---} \\ \text{☷} \end{array} \quad (2402 - 6 = 2396 = 5 \cdot 20^2 + 19 \cdot 20^1 + 16 \cdot 20^0). \quad (1.8)$$

4. RANDOM VARIABLES AND PROCESSES

Numerical analyses in the engineering practice and sciences are performed over **random variables**. Roughly speaking, random variable is such a variable whose value is an outcome from an observation of a **random process**.

A typical random variable is the temperature, i.e., a measure of the average kinetic energy of molecules, whose chaotic moving is a random process. Unlike the temperature, some other physical quantities are not random, but the random process is inherent to their observation. An example is the measured value of a mass. Although the mass of a body (in the non-relativistic physics) is a **constant variable** (Chapter 3), the random process is inherent to its measurement. Values observed by repeated measurements fluctuate randomly about some average¹ so that each observed value of the mass, as well as an average of observed values, has a random component.

Those random variables whose observed values fluctuate about some average may be expressed in the form similar to that in expression (3.2):

$$x = \bar{x} \pm \Delta x \quad @ \quad \text{C.L. (in \%)}, \quad (4.1)$$

but with different interpretation of \bar{x} and Δx . Herein, \bar{x} is either an average of the observed values or just a single observed value. The magnitude of indeterminacy in \bar{x} is quantified by the **random error** (or **uncertainty**) Δx with some trust referred to as a **confidence level** C.L. (in %). Before a detailed

¹ Fluctuations around some average can be observed only if the scale of measuring equipment is precise enough.

explanation of the random error (uncertainty) and its propagation (the subject of Chapter 5) it is necessary to make brief introduction to random processes, probability distributions, probability density, expected values and averages, variances, co-variances and correlations.

4.1 RANDOM PROCESSES

A random (or stochastic) process is such a process which has some indeterminacy in its behaviour. There are various kinds of random processes. An example of a simple random process is illustrated in Fig. 4.1.

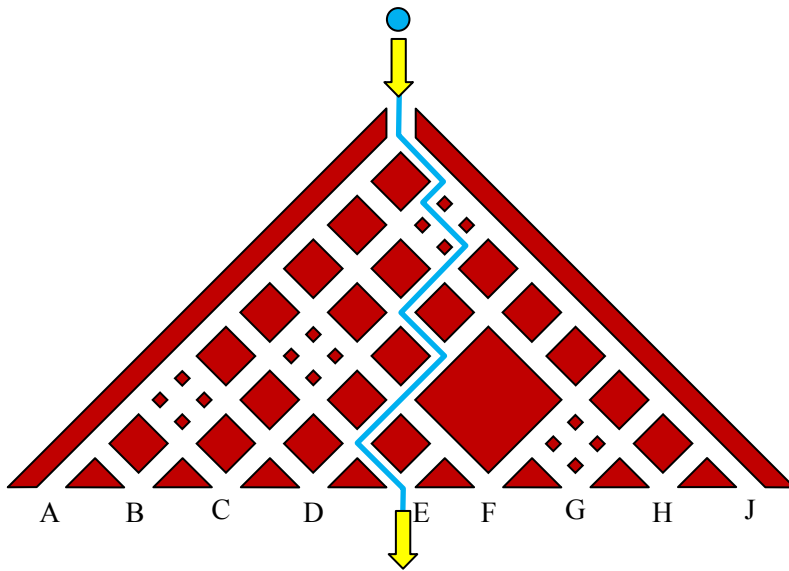


Fig. 4.1. An example of a simple random process

A blue ball is falling down through the labyrinth. At some crossings the ball can continue falling either left or right with equal probability, while at other crossings there is only one possible direction. This means that even when the initial condition (or starting point) is known, there are many (less or more probable) possibilities a random process might go to.

The blue ball leaves the labyrinth at the one of nine exits A , B , C , ..., H or J (in Fig. 4.1 at the exit E), what is then referred to as the one of nine **elementary events** A , B , C , ..., H and J . Events are described in Appendix A.1.

4.1.1 Definition of random variables

A set of all possible outcomes of an observation can consist of *countably many*¹ (finite or infinite) or in *uncountably many*² (infinite) elementary events.

Assigning numbers to elementary events defines random variable. If there are countably many events, the random variable is **discrete**, otherwise it is **continuous**. Since events are occurring randomly, every random variable X can be considered as a quantity that takes its value x randomly from some (discrete or continuous) set of numbers x_k ($k = 1, 2, \dots, n$).

It should be emphasized that random variable in statistics is denoted with italic uppercase letter, while its values as well as constant variables are denoted with italic lowercase letter. On the contrary, in engineering practice and science both types of variables are denoted either with italic lowercase or italic uppercase letters. Most of these variables are random, while other are non-random.

Every system of elementary events can be described either by a scalar random variable or by a **random vector**, whose components are several **independent** random variables. This must be distinguished from **redundant** random variables that are related to each other (see Example 4.1).

Discrete random variables

Events of the countable system of elementary events may be associated with natural numbers from some (finite or infinite) subset of natural numbers, or possibly with real numbers from some countable subset of real numbers. Assigning these numbers to the events of the countable system of elementary events defines a **discrete random variable**, Example 4.1.

Example 4.1. Discrete random variables.

A discrete random variable X in Tab. 4.1 is defined by its values $x \in \{1, 2, \dots, 9\}$ associated with the elementary events A, B, ..., J of the random process described with Fig. 4.1. Another discrete random variable $Y = \sqrt{X}$ in Tab. 4.1 is defined by its values $y \in \{1, \sqrt{2}, \sqrt{3}, \dots, \sqrt{9}\}$ on the same events.

¹ The elements of a **countable set** can be counted one at a time – although the counting may never finish. For example, natural numbers and rational numbers can be counted.

² **Uncountable set** is an infinite set that contains too many elements to be countable. For example, irrational numbers can not be counted even in a finite interval.

Tab. 4.1. Random variables associated to the labyrinth in Fig. 4.1.

Event	A	B	C	D	E	F	G	H	J
X	1	2	3	4	5	6	7	8	9
Y	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{4}$	$\sqrt{5}$	$\sqrt{6}$	$\sqrt{7}$	$\sqrt{8}$	$\sqrt{9}$

Random variables X and $Y = \sqrt{X}$ are strictly dependent, so that only one of them is sufficient; the second one is **redundant**. However, there are many other random variables that can be defined on the given system of events.

The results of the observation given in Fig. 4.1 (blue ball leaves the labyrinth at the exit E) are $x = 5$ and $y = \sqrt{5}$.

Sometimes, events have “coordinates” in a multidimensional space. In such a case, a system of countably many elementary events will be described with a vector whose components are **independent** discrete random variables. A typical example for that is a countable system of elementary events $A_{xy\dots z}$, denoted by indices x, y, \dots, z , which are the possible values of independent random variables X, Y, \dots, Z . These random variables X to Z are then components of a random vector.

Continuous random variables

Elementary events of the uncountable system of events may be associated with real numbers from some uncountable set of real numbers. Assigning these numbers to the elementary events of the uncountable system of events defines a **continuous random variable** in the form of a scalar or a vector.

Typically, a continuous random variable takes all the possible values in a continuous m -dimensional real domain Ω in which it is defined. In the one-dimensional space ($m = 1$), this domain may be either a finite interval $[a, b]$, a semi-infinite interval as $[0, +\infty)$, or the infinite interval $(-\infty, +\infty)$.

Although there are uncountably many possible values in an interval $[a, b]$ (either finite or infinite), only the finitely many rounded values may be observed and then recorded in a finite precision, even when the scale of measuring equipment is analogue (what in theory implies infinite precision). That is, if the used format is a B bit wide binary format, then 2^B different values x_k ($k = 1, 2, 3, \dots, 2^B$) may be recorded. The number of possible values can be very large (e.g. if $B = 32$, then $2^{32} = 4\,294\,967\,296$), but it is still countable and finite, Fig. 4.2.

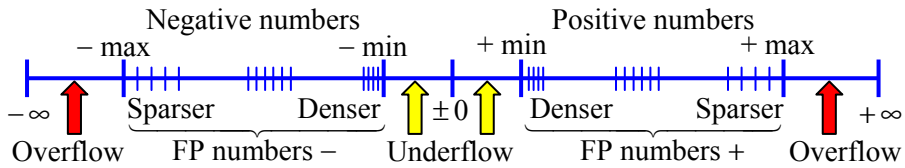


Fig. 4.2 Ranges in the floating-point formats

If recorded value x_k were rounded by truncation, then each of them would represent the finite interval $\Omega_k = [x_k, x_{k+1})$ ($k \in \{1, 2, 3, \dots, 2^B - 1\}$). If the lowest possible value x_1 denotes negative overflow ($x_1 = -\infty$) and if the highest possible value x_{2^B} denotes positive overflow ($x_{2^B} = +\infty$), then the complete set of all possible intervals $[x_k, x_{k+1})$ ($k = 1, 2, 3, \dots, 2^B - 1$) builds the continuous and infinite interval $(-\infty, \infty)$.

Although the values of continuous random variables cannot be used as indices in denoting elementary events (like discrete random variables), they can be used in describing uncountably many events by using a logical statement which includes inequalities with random variables and their values, Example 4.2.

Example 4.2. Continuous random variables.

The lifetime T of light-bulbs is a continuous random variable. An elementary event $T = t$ that the lifetime T of randomly selected light-bulb is exactly t is **unobservable**, because a real number can have an infinite number of nonzero decimal places. The scale of any measuring equipment has finite precision so that an event $t_1 \leq T < t_2$ is observable when the observed lifetime T of randomly selected light-bulb takes a value within the interval $[t_1, t_2)$.

Events $t_k \leq T < t_{k+1}$ ($k = 1, 2, \dots$), starting with $t_1 = 0$, build a complete system of countably many elementary events in the semi-infinite interval $[0, \infty)$.

4.1.2 Quantification of random processes

There are several methods to quantify a random process. Some of them, like

- studying minimal population, or
- studying equivalent non-random process,

provide confident information about a random process, while in other, like

- sampling,

there is only a hope that conclusions made about a random process are valid enough.

Population in statistics is an entire group (or collection) of entities (people, animals, things, etc.) or events that have something in common and about which some descriptions or conclusions are made. Examples of populations are: members of a club, all textbooks published last year by some publisher, all possible paths in the labyrinth, etc.

More widely, population is every set of data that consists of all conceivably possible (or hypothetically possible) observations of a given phenomenon.

Minimal population is a part of whole population with a minimal number of members required to quantify a random process. For example, the minimal population in the simple random process in Fig. 4.3a consists of all possible paths (eight of them), Fig. 4.3b.

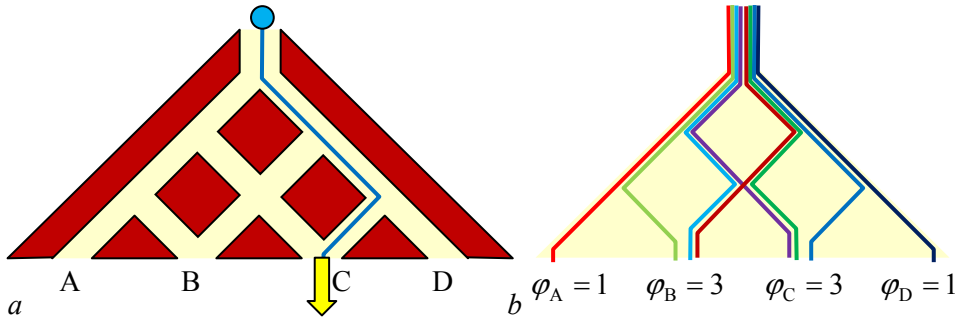


Fig. 4.3. Simple labyrinth, *a* – single path, *b* – population of paths

A whole population may consist of a lot of equal minimal populations. A collection of several equal minimal populations quantify the same random process as only one of them.

Population frequency

Minimal population for the labyrinth in Fig. 4.3a consists of $n = 8$ paths (Fig. 4.3b) that are ending at the exits A, B, C and D respectively in the quantities

$$\varphi_A = 1, \varphi_B = 3, \varphi_C = 3 \text{ and } \varphi_D = 1. \quad (4.2)$$

Quantities $\varphi_A, \varphi_B, \dots$ are called *population frequencies*. It is obvious that

$$0 < \varphi_A, \varphi_B, \dots < n \text{ and } \varphi_A + \varphi_B + \dots = n. \quad (4.3)$$

Population frequencies are proportional to the number of paths n .

Probability

If paths in Fig. 4.3b occur with the equal probability $P(\text{path}) = 1/n$, then the probability $P(A)$ (Appendix A.2) that an elementary event A will occur is the

quotient

$$P(A) = \varphi_A P(\text{path}) = \varphi_A / n, \quad (4.4)$$

where $n = \varphi_A + \varphi_B + \dots$ is the sum of the population frequencies $\varphi_A, \varphi_B, \dots$ for all events in the complete set of elementary events $\{A, B, \dots\}$.

In the given example (Fig. 4.3), $n = 8$, $P(\text{path}) = 0,125$ so that

$$P(A) = 0,125, \quad P(B) = 0,375, \quad P(C) = 0,375 \quad \text{and} \quad P(D) = 0,125. \quad (4.5)$$

Unfortunately, random processes are not always that simple so that quantification of random processes by studying their minimal populations can not be applied to all of them.

4.1.3 Equivalent non-random process

A possible way to quantify a random process is to simulate it by an equivalent non-random process. Consider that some quantity (e.g. 1024) of equal balls¹ starts simultaneously falling down through the labyrinth in Fig. 4.4.

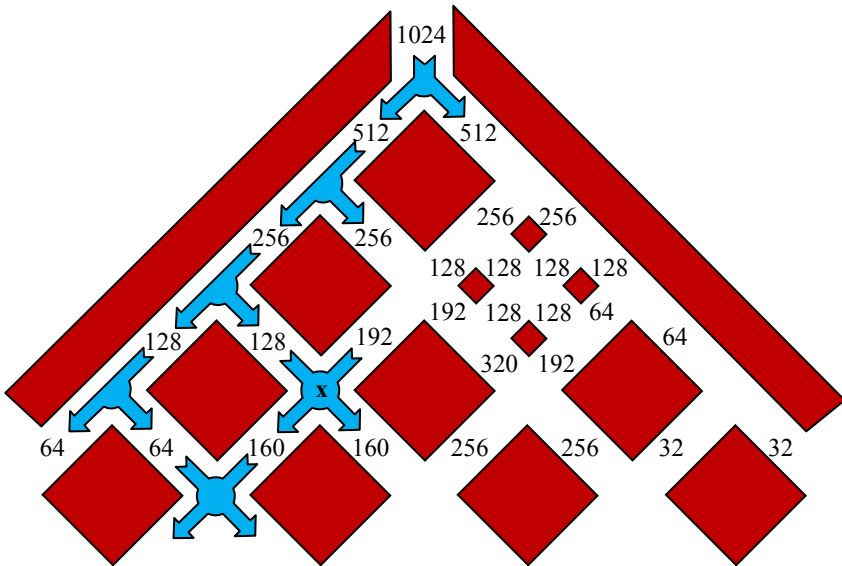


Fig. 4.4. Streams of balls in the upper part of labyrinth in Fig. 4.1

¹ Although the balls are initially equal, after passing through the labyrinth, each ball will have a path associated to it. Thus, after leaving the labyrinth, the balls will differ by their history. However, some balls can share the same history so that they can be equal.

At those crossings where the single ball can continue falling either left or right with equal probability, the incoming stream of balls is divided into two equal streams: one going left and another going right. In this way, the random process in the labyrinth can be simulated by the equivalent non-random process.

Consider now the non-random process equivalent to the random process in Fig. 4.1, in which all balls start falling down through the labyrinth, as it is illustrated in Fig. 4.4. Each ball can pass up to 10 crossings before it reaches an exit so that the stream of balls will be cut in two up to 10 times. Since a single ball cannot be divided, the results of those dividings must be integers greater or equal to the number 1. In the given example, this can be achieved by the minimal quantity of $2^{10} = 1024$ balls.

Numbers between crossings in Fig. 4.4 represent the quantity of passed balls. At the entrance of the labyrinth the $n = 1024$ balls are divided into two streams of 512 balls. Each stream is then divided into two sub-streams of 256 balls, etc. However, when two streams come to the same crossing, they are first merged and then divided. For example, two incoming streams of $128 + 192 = 320$ balls are merged at the crossing marked with “x” and then divided into two equal streams each containing $320/2 = 160$ balls. Finally, the balls are leaving the observed labyrinth at the exits (events A, B, C, ..., H and J) in the quantities that are equal to the population frequencies $\varphi_A, \varphi_B, \dots, \varphi_H, \varphi_J$ (Tab. 4.2).

Probability

Probability $P(A)$ that an elementary event A will occur is the quotient

$$P(A) = \varphi_A / n, \quad (4.6)$$

where $n = \varphi_A + \varphi_B + \dots$ is the sum of the population frequencies $\varphi_A, \varphi_B, \dots$ for all events in the complete set of elementary events $\{A, B, \dots\}$, Example 4.3.

Example 4.3. Some properties of probability.

Probabilities that a blue ball in the random process depicted in Fig. 4.1 will leave the labyrinth at the corresponding exit (elementary events A, B, ..., J) are

$$P(A) = \varphi_A / n, \quad P(B) = \varphi_B / n, \quad \dots, \quad P(J) = \varphi_J / n. \quad (4.7)$$

The obtained probabilities are listed in Tab. 4.2.

6. REGRESSION

Regression consists in fitting a function to a set of data points by finding such parameters in the function that provide the best fit. The fitted function is referred to as a *regression model*, while their graphical representation is referred to as a *trendline*. Regarding the applied function, the most popular regressions are

- Linear regression
- Polynomial regression
- Exponential regression
- Logarithmic regression
- Power regression

Other methods of fitting a function to a set of data points are described in Volume III and Volume IV of Numerical Methods.

There are two approaches in performing regressions: the first approach when the variance in residual of a function is minimised, and the second approach when the variance in residuals of all variables is minimised. Regressions based on the first approach may be referred to as the “simple” so as to distinguish them from regressions based on the second approach, which are referred to as the “orthogonal”.

6.1 LINEAR REGRESSION

The most popular types of linear regressions are

- Simple linear regression
- Linear regression through a fixed point
- Orthogonal linear regression

6.1.1 Simple linear regression

In the regression model, the linear function

$$F(x) = a + bx \quad (6.1)$$

is fitted to a set of N data points (x, y) by finding those values of constants a and b that minimise the variance in residual. The residual ε_y in each data point (x, y)

$$\varepsilon_y = y - F(x) \quad (6.2)$$

is the difference between the value y and the value $F(x)$ predicted by the regression model for the given x . It represents a "vertical" distance between the observed data point and the fitted curve. This is illustrated in Fig. 6.1.

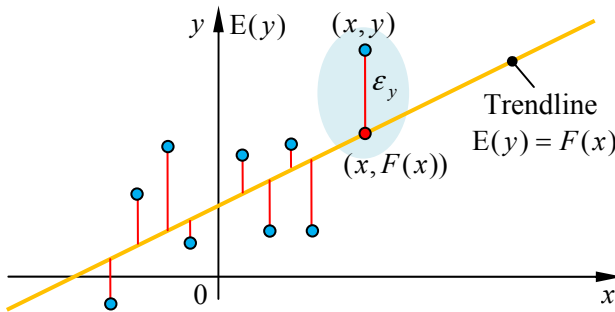


Fig. 6.1. Residual in simple linear regression

To explain the meaning of the residual ε_y , the function $F(x)$ should be eliminated from the expressions (6.1) and (6.2), giving

$$y = a + bx + \varepsilon_y. \quad (6.3)$$

If x and y are both constant variables, then the residual ε_y may be

- the nonlinear component of the true function $y(x)$, which is excluded from the linear regression model, or
- the function $\varepsilon_y = \varepsilon_y(z_1, z_2, \dots)$ of independent variables z_1, z_2, \dots , whose influence to y is unknown, whether or no the function $\varepsilon_y(z_1, z_2, \dots)$ be unknown or variables z_1, z_2, \dots are not observed together with x and y .

In both cases, the regression model $F(x) = a + bx$ is just an approximation of the true function $y(x)$.

If y is just a random variable, then the residual ε_y is merely an observation error in y , while $F(x)$ predicts the expected value $E(y)$ of y for the given x . The

argument x in $F(x)$ is assumed to be a constant variable, otherwise orthogonal linear regression should be performed (Chapter 6.1.3).

It is convenient to introduce notation for averages in x and y :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x(i), \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y(i), \quad (6.4)$$

and for average residual

$$\bar{\varepsilon}_y = \frac{1}{N} \sum_{i=1}^N \varepsilon_y(i) = \frac{1}{N} \sum_{i=1}^N [y(i) - F(x(i))] = \bar{y} - (a + b\bar{x}). \quad (6.5)$$

The sample variance in residual

$$s_\varepsilon^2 = \frac{1}{N-1} \sum_{i=1}^N (\varepsilon(i) - \bar{\varepsilon})^2 = s_y^2 - 2bs_{xy} + b^2s_x^2, \quad (6.6)$$

where s_x and s_y (4.72) are sample variances in x and y , while s_{xy} (4.52) is their covariance.

Two unknown constant coefficients a and b can be determined by the conditions

$$\bar{\varepsilon} = 0, \quad \partial s_\varepsilon^2 / \partial b = 0, \quad (6.7)$$

that minimises both: the magnitude of average residual ($\min |\bar{\varepsilon}| = 0$) and the empirical variance s_ε^2 . These two conditions give

$$a = \bar{y} - b\bar{x} \quad \text{and} \quad b = s_{xy} / s_x^2. \quad (6.8)$$

The final form of the regression model

$$F(x) = \bar{y} + (x - \bar{x}) s_{xy} / s_x^2. \quad (6.9)$$

Therefore, a trendline in the simple linear regression is the straight line passing through the point (\bar{x}, \bar{y}) with the slope s_{xy} / s_x^2 .

The minimal sample variance in residual for given b has a value

$$\min s_\varepsilon = \sqrt{(s_y - s_{xy}/s_x)(s_y + s_{xy}/s_x)} = s_y \sqrt{(1-r)(1+r)}, \quad (6.10)$$

where $r = s_{xy} / s_x s_y$ (4.88) is the sample coefficient of regression.

Simple linear regression is illustrated with Example 6.1. The practical application is supported by Algorithm 6.1.

Example 6.1. Simple linear regression.

Simple linear regression is performed using data in Tab. 6.1. Results are illustrated by diagrams in Fig. 6.2.

Tab. 6.1. Data for linear regression (example)

i	1	2	3	4	5	6	7
$x(i)$	0,1	0,2	0,4	0,5	0,6	0,8	0,95
$y(i)$	1,115	1,115	1,38	1,6	1,5	1,82	1,93

Average values and variances are

$$\begin{aligned}\bar{x} &= 0,507142857, \quad \bar{y} = 1,494285714, \\ s_x^2 &= 0,093690476, \quad s_y^2 = 0,101320238, \quad s_{xy} = 0,095214286.\end{aligned}\quad (6.11)$$

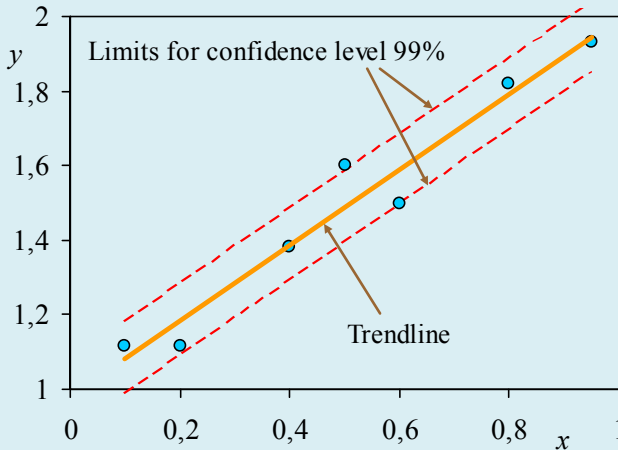
Constants

$$b = s_{xy} / s_x^2 = 1,0163, \quad a = \bar{y} - b\bar{x} = 0,9789. \quad (6.12)$$

The regression model (trendline) is determined by the function

$$F(x) = 0,9789 + 1,0163x. \quad (6.13)$$

The squared correlation coefficient $r^2 = s_{xy}^2 / (s_x^2 s_y^2) = 0,955$ (Chapter 4.2.7).

**Fig. 6.2.** Example of simple linear regression

Sample variance and standard deviation in residual are

$$s_\varepsilon^2 = s_y^2 - 2bs_{xy} + b^2s_x^2 = 0,00456 \quad \text{and} \quad s_\varepsilon = 0,0675. \quad (6.14)$$

Uncertainties in residual $\Delta\epsilon_y = t_{M\alpha}s_\epsilon/\sqrt{N}$ (5.22) for several confidence levels (Chapter 5.3), $N = 7$ and $M = N - 1$ are given in Tab. 6.2.

Tab. 6.2. Uncertainties in residual (example)

$1-\alpha$	50%	70%	90%	95%	99%
$t_{M\alpha}$	0,718	1,134	1,943	2,447	3,707
$\Delta\epsilon_y$	0,0183	0,0289	0,0495	0,0624	0,0945

Algorithm 6.1. Simple linear regression

Input data: $x(i)$, $y(i)$, N

Output data: a , b , r^2

// Averages

$avex = avey = 0$

for $i = 1$ to N

$avex = avex + x(i)$

$avey = avey + y(i)$

endfor

$avex = avex / N$

$avey = avey / N$

$$// \bar{x} = \frac{1}{N} \sum_{i=1}^N x(i)$$

$$// \bar{y} = \frac{1}{N} \sum_{i=1}^N y(i)$$

// Sample variances and covariance

$sx2 = sy2 = sxy = 0$

for $i = 1$ to N

$sx2 = sx2 + (x(i) - avex)^2$

$$// s_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x(i) - \bar{x})^2$$

$sy2 = sy2 + (y(i) - avey)^2$

$sxy = sxy + (x(i) - avex)(y(i) - avey)$

$$// s_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y(i) - \bar{y})^2$$

endfor

$sx2 = sx2 / (N - 1)$

$sy2 = sy2 / (N - 1)$

$$// s_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x(i) - \bar{x})(y(i) - \bar{y})$$

$sxy = sxy / (N - 1)$

// Coefficients

$b = sxy / sx2$, $a = avey - b \cdot avex$

$$// b = s_{xy} / s_x^2$$

$$// a = \bar{y} - b\bar{x}$$

$r^2 = sxy \cdot sxy / (sx2 \cdot sy2)$

$$// r^2 = s_{xy}^2 / s_x^2 s_y^2$$

end

Boris Obsieger
NUMERICAL METHODS I
Basis and Fundamentals

The series of books *Numerical Methods* is written primarily for students at technical universities, but also as a useful handbook for engineers, PhD students and scientists.

This volume introduces the reader into numerical systems and representation of numbers in digital computers. Possibly the most important part of this book are descriptions of differences between constant and random variables, related types of errors and error propagations. These topics are supplemented with various types of regression analyses. Finally, direct and iterative methods for finding roots of polynomials are explained.

Practical application is supported by 76 examples and 13 algorithms. For reasons of simplicity, algorithms are written in pseudo-code, so they can easily be included in any computer program.

About author

Boris Obsieger, D.Sc., professor at the University of Rijeka, Croatia. Head of Section for Machine Elements at the Faculty of Engineering in Rijeka. Holds lectures on Machine Elements Design, Robot Elements Design, Numerical Methods in Design and Boundary Element Method. Several invited lectures. President of CADAM Conferences. Main editor of international journal *Advanced Engineering*. Author of several books and a lot of scientific papers.



Standard option

Printed in **colour**

ISBN 978-953-6326-66-2

ISBN 978-953-57117-1-1

eBook in **colour**

ISBN 978-953-7919-25-2

Adobe® Digital Editions
by Ingram Digital