Lecture notes on stochastics – Part 1

Rome, 3-4 May 2018

A brief introduction to probability and stochastics

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens, Greece
(dk@itia.ntua.gr, http://www.itia.ntua.gr/dk/)

Presentation available online: http://www.itia.ntua.gr/1835/

Senza probabilità (Without Probability): An example

- Problem: study the storage and outflow of a (toy) water supply reservoir in discrete time with ridiculously simple assumptions.
- Assumption 1 – **the ideal hydrological model**: The inflow I to the reservoir is constant for any time step, equal to 10 units.
- Assumption 2 – **the perfect socio-hydrological model**: If there is plenty of water in the reservoir, people consume more, while the consumption is reduced when the storage is low. We assume that this behaviour is expressed precisely by an exponential function: $Q = \varphi(S) = 0.2\,\mathrm{e}^{0.3\,S}$, where Q is the outflow and S the storage.
- Discrete time dynamics $Q_i = \varphi(S_{i-1})$, $S_i = S_{i-1} + I - Q_i$
- Question 1: Assume a specific initial storage S_0 in the interval (5, 15) and find S_1.
- Question 2: With the same initial condition, find S_{50}.
- Question 3: Is the system dynamics deterministic or stochastic?
- Question 4: Is the **system predictable (i.e., deterministic) or unpredictable (i.e., stochastic, random)?**

Impacts of the creation of a single electron somewhere at the limit of the known universe

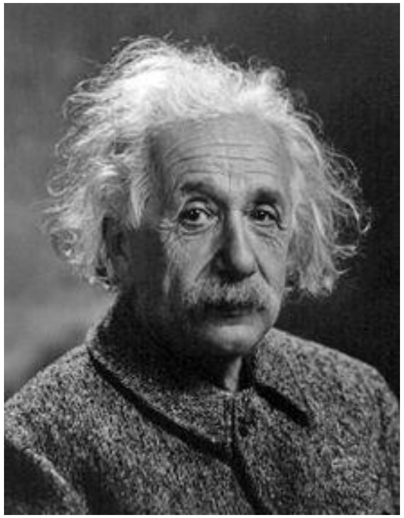
Perturbation scale	Cause	Effect	Time frame
1. Microscopic	An electron has been created at a distance of 10^{10} light years from Earth and has exerted gravitational attraction on the molecules of the atmosphere	After 50 collisions the trajectories of molecules would have changed (different molecules would collide)	10 ns (justifiably assuming $\sim 10^{10}$ collisions per second)
2. Macroscopic	Turbulence	Change in the fine structure of turbulence	1 min
3. Local	Turbulence	Change in the large (km) scale atmospheric turbulence (wind, clouds)	1 h – 1 d
4. Global		Change in the general circulation of the atmosphere (depressions, fronts→ a storm that would not occur without that electron)	1-2 weeks

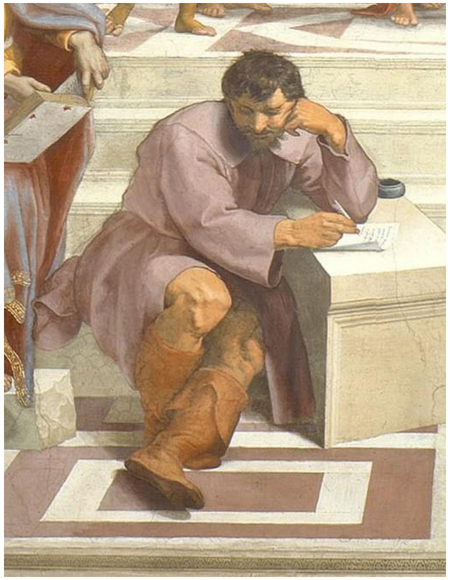
Adapted from Ruelle (1979, 1991, p. 75); based on Berry (1978) and some ideas of E. Borel και B. V. Chirikov.

So, who does not play dice?

Jedenfalls bin ich überzeugt, daß der nicht würfelt
I, at any rate, am convinced that He [God] does not throw dice

(Albert Einstein, in a letter to Max Born in 1926)



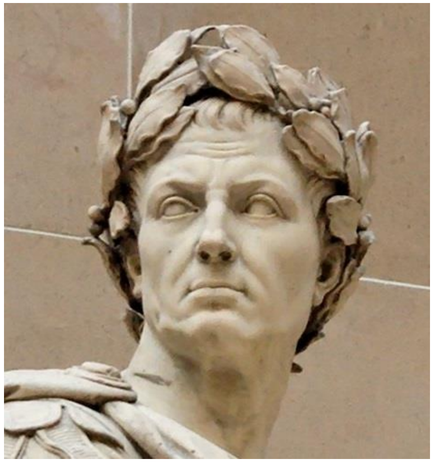


Αἰών παῖς ἐστι παίζων πεσσεύων
Time is a child playing, throwing dice

(Heraclitus; ca. 540-480 BC; Fragment 52)

Ἀνερρίφθω κύβος	Iacta alea est
Let the die have been cast	The die has been cast
[Plutarch’s version, in Greek]	[Suetonius’s version, in Latin]

(Julius Caesar, 49 BC, when crossing Rubicon River)

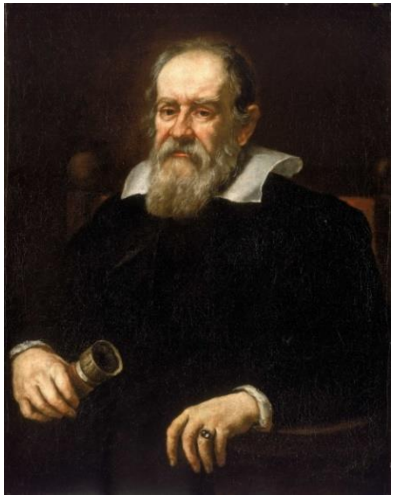


From old times dice games fascinated people—but perhaps not scientists and teachers

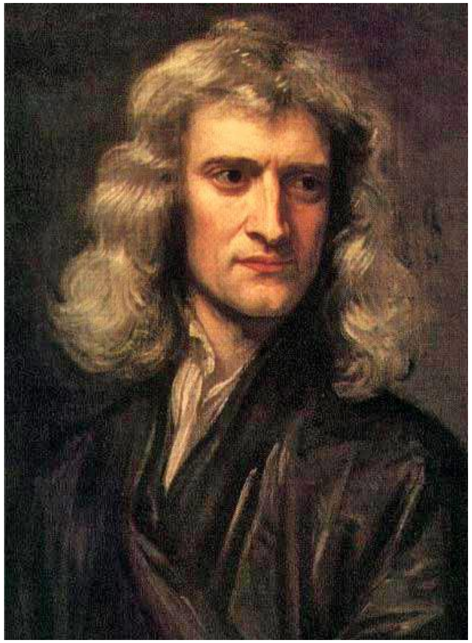
- All these dice are of the period 580-570 BC from Greek archaeological sites:
  - Left, Kerameikos Ancient Cemetery Museum, Athens, photo by author
  - Middle: Bronze die (1.6 cm), Greek National Archaeological Museum, www.namuseum.gr/object-month/2011/apr/7515.png
  - Right: Terracotta die (4 cm) from Sounion, Greek National Archaeological Museum, http://www.namuseum.gr/object-month/2011/dec/dies_b.png
- Much older dice (up to 5000 years old) have been found in Asia (Iran, India).

Modern determinism and the clockwise universe

- Johannes Kepler (1571-1630), Galileo Galilei (1564-1642) and René Descartes (1596-1650) introduced mathematical concepts to natural philosophy (science).
- They also introduced the idea of a clockwork universe, leading to the philosophical proposition of *determinism*, still widely accepted in science.
- Determinism was perfected by the French mathematician and astronomer Pierre-Simon Laplace (1749-1827; cf. *Laplace's demon,* a hypothetical entity that knows the precise location and momentum of every atom in the universe at present, and can deduce the future and the past using Newton's laws.).
- According to deterministic thinking, the roots of uncertainty about future are subjective, i.e. rely on the fact that we do not know exactly the present, or we do not have good enough methods and models. It is then a matter of time to eliminate uncertainty, with better data and better models.



Newton’s awareness of the fragility of the universe (rejection of determinism)

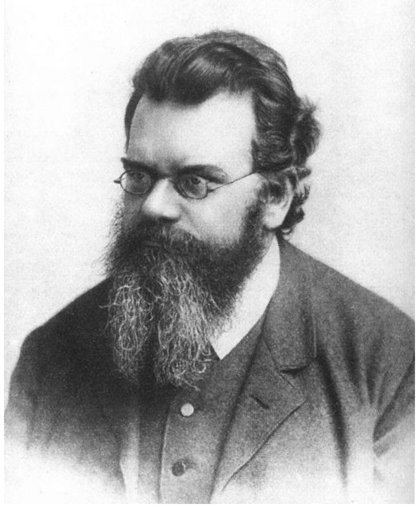


“For while comets move in very eccentric orbs in all manner of positions, blind fate could never make all the planets move one and the same way in orbs concentric, some inconsiderable irregularities excepted which may have arisen from the mutual actions of comets and planets on one another, and which will be apt to increase, till this system wants a reformation” (Newton, *Opticks*, Query 31).

- Newton regarded the complexity and fragility of the universe as proof of the existence of God.
- He rejected Leibniz’ thesis that God would necessarily make a perfect world which requires no intervention from the creator.
- Newton simultaneously made an argument from design and for the necessity of intervention.

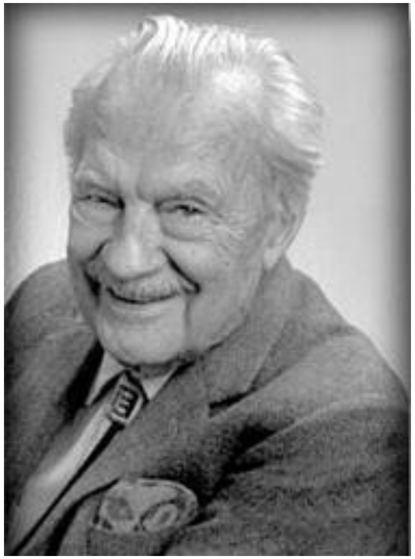
From the almighty determinism of the 17th century to the probabilistic world of the 20th century

- **Statistical physics** (cf. Boltzmann) used the probabilistic concept of entropy (which is nothing other than a quantified measure of uncertainty defined within the probability theory; see below) to explain fundamental physical laws (most notably the Second Law of Thermodynamics), thus leading to a new understanding of natural behaviours and to powerful predictions of macroscopic phenomena.
- **Dynamical systems** theory (cf. Poincare) has shown that uncertainty can emerge even from pure, simple and fully known deterministic (chaotic) dynamics, and cannot be eliminated.
- **Quantum theory** (cf. Heisenberg) has emphasized the intrinsic character of uncertainty and the necessity of probability in the description of nature.



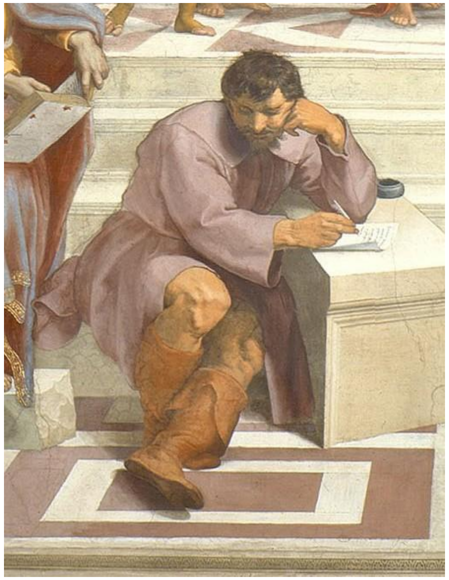
From the almighty determinism of the 17th century to the probabilistic world of the 20th century (2)

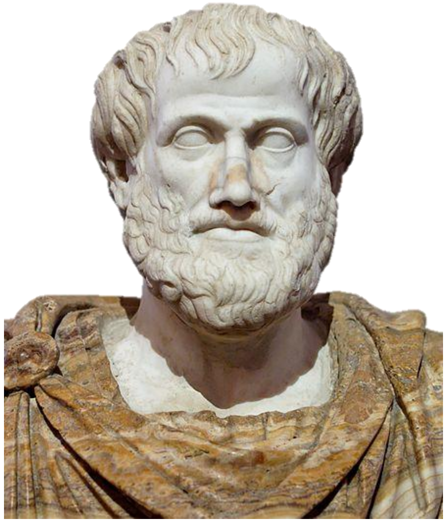
- Developments in **mathematical logic**, and particularly **Gödel's incompleteness theorem**, challenged the almightiness of deduction (inference by mathematical proof). Ironically, **Kurt Gödel** anticipated by one day (in 1930) **David Hilbert** who pronounced the opposite with his famous aphorism (also inscribed in his tombstone at Göttingen) "*Wir müssen wissen, wir werden wissen*" ("***We must know, we will know***").
- Developments in **numerical mathematics** (cf. **Nicholas Metropolis**) highlighted the effectiveness of stochastic methods in solving even purely deterministic problems, such as **numerical integration** in high-dimensional spaces and **global optimization** of non-convex functions (where stochastic techniques, e.g. evolutionary algorithms or simulated annealing, are in effect the only feasible solution in complex problems that involve many local optima).
- Advances in **evolutionary biology** emphasize the importance of stochasticity (e.g. in selection and mutation procedures and in environmental changes) as a driver of evolution.

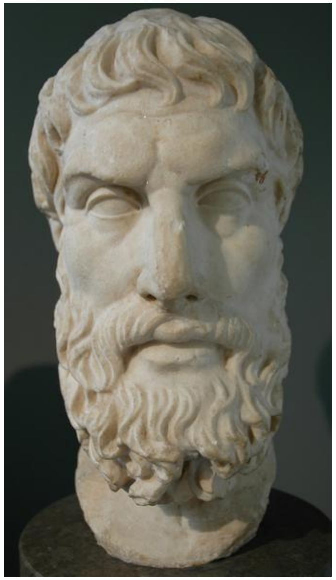


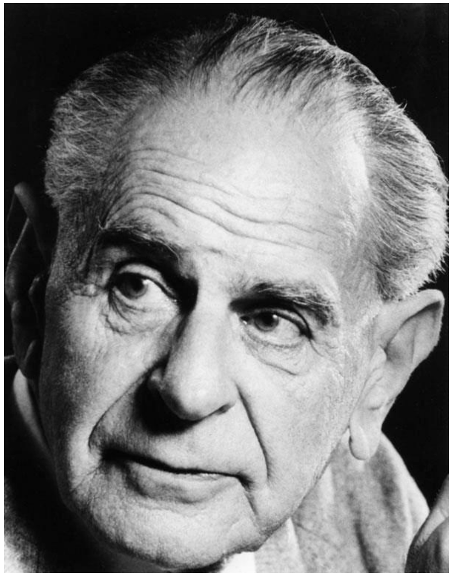
Indeterminism vs. determinism

- In *indeterminism*, a philosophical belief contradictory to determinism, uncertainty may be a structural element of nature and thus cannot be eliminated.
- Indeterminism has its origin in the Greek philosophers **Heraclitus** (*ca*. 535–475 BC), **Aristotle** (384 – 322 BC) and **Epicurus** (341–270 BC).
- Its relationship with modern science was theorized by the Austrian-British philosopher **Karl Popper** (1902-1994).
- In science, indeterminism largely relies on the notion of ***probability***, which according to **Popper** is the extension (quantification) of the **Aristotelian** idea of ***potentia*** (***dynamis***). Practically, the idea is that several outcomes can be produced by a specified cause, while in deterministic thinking only one outcome is possible (albeit difficult to predict which one).









The meaning of probability (by examples)

(1) A fair coin has a probability of 0.5 of heads, and likewise 0.5 of tails; so the probability of tossing two heads in a row is 0.25.

(2) There is a 10% probability of rain tomorrow.

(3) There is a 10% probability of rain tomorrow according to the weather forecast.

(4) Fortunately there is only a 5% probability that her tumor is malignant, but this will not be known for certain until the surgery is done next week.

(5) Smith has a greater probability of winning the election than does Jones.

(6) I believe that there is a 75% probability that she will want to go out for dinner tonight.

(7) I left my umbrella at home today because the forecast called for only a 1% probability of rain.

(8) Among 100 patients in a clinical trial given drug *A*, 83 recovered, whereas among 100 other patients given drug *B*, only 11 recovered; so new patients will have a higher probability of recovery if treated with drug A.

Source of examples: Gauch (2003).

The utility of probability

- Commonly, probability is regarded to be a branch of applied mathematics that provides tools for data analysis (and also theorizes games of chance).
- Historically, as modern science was initiated from deterministic views of the world, probability had a marginal role for peculiar unpredictable phenomena.
- Nonetheless, probability is a more general concept that helps shape a consistent, realistic and powerful view of the world.
- Probability has provided grounds for philosophical concepts such as indeterminism and causality, as well as for extending the typical mathematical logic, offering the mathematical foundation of induction.
- In typical scientific and technological applications, probability provides the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions.

See more details in Koutsoyiannis (2008).

Deduction and induction

- In mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*). This type of reasoning consists of repeated application of strong syllogisms such as:

  If A is true, then B is true;
  A is true;
  Therefore, B is true.

  If A is true, then B is true;
  B is false;
  Therefore, A is false.

- Deduction uses a set of axioms to prove propositions known as theorems, which, given the axioms, are irrefutable, absolutely true statements. It is also irrefutable that deduction is the preferred route to truth; the question is, however, whether or not it has any limits.
- David Hilbert's belief "*Wir müssen wissen, wir werden wissen*", more formally known as *completeness*, according to which any mathematical statement could be proved or disproved by deduction from axioms, has been proved to be invalid.
- In everyday life, however, we use weaker syllogisms of the type:

  If A is true, then B is true;
  B is true;
  Therefore, A becomes more plausible.

  If A is true, then B is true;
  A is false;
  Therefore, B becomes less plausible.

- The latter type of syllogism is called *induction* (the Aristotelian *epagoge*). It does not offer a proof that a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible.
- An important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data.

Definition of probability

- According to Kolmogorov's (1933) axiomatization, probability theory is based on three fundamental concepts and four axioms.
- The concepts, i.e., the triplet (Ω, Σ, P) called *probability space*, are:
  1. A non-empty set Ω, sometimes called the *basic set*, *sample space* or the *certain event* whose elements ω are known as *outcomes* or *states*.
  2. A set Σ known as *σ-algebra* or *σ-field* whose elements E are subsets of Ω, known as *events*. Ω and $\emptyset$ are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
  3. A function P called *probability* that maps events to real numbers, assigning each event E (member of Σ) a number between 0 and 1.
- The four axioms, which define the properties of P, are:
  - I. **Non-negativity**: For any event A, $P(A) \geq 0$.
  - II. **Normalization**: $P(\Omega) = 1$.
  - III. **Additivity**: For any events A, B with $AB = \emptyset$, $P(A + B) = P(A) + P(B)$.
  - IV. **Continuity at zero**: If $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ is a decreasing sequence of events, with $A_1 A_2 \ldots A_n \ldots = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

    [Note: In the case that Σ is finite, axiom IV follows from axioms I-III; in the general case, however, it should be put as an independent axiom.]

The concept of a random variable

- A random variable $\underline{x}$ is a function that maps outcomes to numbers, i.e. quantifies the sample space Ω.
- More formally, a real single-valued function $\underline{x}(\omega)$, defined on the basic set Ω, is called a *random variable* if for each choice of a real number a the set $\{\underline{x} < a\}$ for all ω for which the inequality $\underline{x}(\omega) < \alpha$ holds true, belongs to Σ.
- With the notion of the random variable we can conveniently express events using basic mathematics. In most cases this is done almost automatically. For instance a random variable $\underline{x}$ that takes values 1 to 6 is intuitively assumed when we deal with a die through.
- We must be attentive that a random variable is not a number but a function. Intuitively, we could think of a random variable as an object that represents simultaneously all possible outcomes and only them.
- A particular value that a random variable may take in a random experiment, else known as a *realization* of the variable, is a number.
- We can denote a random variable by an underlined letter, e.g. $\underline{x}$ and its realization with a non-underlined letter x (another convention is to use an upper case letter, e.g. X, for the random variable and a lower case letter, e.g. x, for its realization. In any case, random variables and values thereof two should not be confused).

Probability distribution function

- *Distribution function* is a function of the real variable x defined by

  $$F(x) := P\{\underline{x} \le x\}$$

  where $\underline{x}$ is a random variable.
- The random variable with which this function is associated is not an argument of the function. If there risk of confusion (e.g. there are many random variables), the random variable is usually denoted as a subscript (e.g. $F_{\underline{x}}(x)$). Typically $F(x)$ has a mathematical expression depending on some parameters. The domain of $F(x)$ is not identical to the range of the random variable $\underline{x}$; rather it is always the set of real numbers.
- The distribution function is a non-decreasing function obeying the relationship

  $$0 = F(-\infty) \le F(x) \le F(+\infty) = 1$$
- For its non-decreasing attitude, in the English literature the distribution function is also known as *cumulative distribution function* (cdf) – though “cumulative” is not necessary. In practical applications the distribution function is also known as *non-exceedence probability*. Likewise, the non-increasing function

  $$\overline{F}(x) = P\{\underline{x} > x\} = 1 - F(x)$$

  is known as *exceedence probability* (or survival function, survivor function, tail function).
- The distribution function is always continuous on the right; however, if the basic set Ω is finite or countable, $F(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant between them (staircase-like). Such random variable is called *discrete*. If $F(x)$ is a continuous function, then the random variable is called *continuous*. A *mixed* case is also possible; in this the distribution function has some discontinuities on the left, but is not staircase-like.
- For continuous random variables, the inverse function $F^{-1}(\ )$ of $F(\ )$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x, called *u-quantile* of the variable $\underline{x}$, that is:

  $$x_u = F^{-1}(u)$$

Probability density (or mass) function

- In continuous variables any particular value *x* has zero probability to occur. However, we can still tell which of two outcomes is more probable by examining the ratio of the two probabilities. As this is a $0/0$ expression, having in mind l'Hôpital's rule, we need to examine the ratio of derivatives of probabilities.
- The derivative of the distribution function is called the *probability density function*:

$$f(x) \coloneqq \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

- The basic properties of $f(x)$ are

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

- Obviously, the probability density function does not represent a probability; therefore it can take values higher than 1. Its relationship with probability is described by the following equation:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leq \underline{x} \leq x + \Delta x\}}{\Delta x}$$

- The distribution function can be calculated from the density function by

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y$$

- In discrete random variables, the density is a sequence of Dirac δ functions. It is thus more convenient to use the so-called *probability mass function* $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, $j = 1,\ldots,w$, where *w* is the number of possible outcomes (which can be infinite).

Some common distributions

Name	Probability density function	Distribution function
Uniform in [0, 1]	$f(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$F(x) = \max(0, \min(x, 1))$
Exponential	$f(x) = \begin{cases} \mathrm{e}^{-x/\mu} / \mu & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$	$F(x) = \begin{cases} 1 - \mathrm{e}^{-x/\mu} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$
Normal	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$	$F(x) =$ $\dfrac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\dfrac{(u-\mu)^2}{2\sigma^2}\right) \mathrm{d}u$

Independent and dependent events, conditional probability

- Two events A and B are called *independent* (or *stochastically independent*), if

$$P(AB) = P(A)P(B)$$

- Otherwise A and B are called (*stochastically*) *dependent*.
- The definition can be extended to many events. Thus, the events A_1, A_2, …, are *independent* if for any finite set of distinct indices i_1, i_2, …, i_n:

$$P\left(A_{i_1} A_{i_2} \dots A_{i_n}\right) = P\left(A_{i_1}\right) P\left(A_{i_2}\right) \dots P\left(A_{i_n}\right)$$

- The handling of probabilities of independent events is thus easy. However, this is a special case because usually natural events are dependent. In the handling of dependent events the notion of *conditional probability* is vital.
- By definition (Kolmogorov, 1933), conditional probability of the event A given B (i.e. under the condition that the event B has occurred) is the quotient

$$P(A|B) \coloneqq \frac{P(AB)}{P(B)}$$

- Obviously, if $P(B) = 0$, this conditional probability cannot be defined, while for independent A and B, $P(A|B) = P(A)$. It follows that

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

- From this it follows the *Bayes theorem*:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)}$$

Random number generation

- **Sequence of random numbers** is a sequence of numbers x_i whose every one statistical property is consistent with that of a sample from a sequence of independent identically distributed random variables $\underline{x}_i$ (adapted from Papoulis, 1990).
- **Random number generator** is a device (typically computer algorithm) which generates a sequence of random numbers x_i with given distribution $F(x)$. As most algorithms are purely deterministic, sometimes the numbers are called pseudorandom—but this in not necessary.
- Random number generation is also known as **Monte Carlo** sampling.
- The basis of practically all random generators is the uniform distribution in [0,1]. A typical procedure is the following:
  - We generate a sequence of integers q_i from the recursive algorithm

    $$q_i = (k\, q_{i-1} + c) \bmod m$$

    where k, c and m are appropriate integers (e.g. $k = 69\,069$, $c = 1$, $m = 2^{32} = 4\,294\,967\,296$ or $k = 7^5 = 16\,807$, $c = 0$, $m = 2^{31} - 1 = 2\,147\,483\,647$; Ripley, 1987, p. 39).
  - We calculate the sequence of random numbers u_i with uniform distribution in [0,1] by

    $$u_i = q_i / m$$

- A more recent and better algorithm is the so-called *Mersenne twister* (en.wikipedia.org/wiki/Mersenne_twister). It is available in most languages and software packages. For example, for Excel (which by default includes the function rand) the Mersenne twister algorithm, called NtRand, can be found in www.ntrand.com/download/.
- A direct (but sometimes time demanding) algorithm to produce random numbers x_i from *any* $F(x)$ given random numbers u_i with uniform distribution in [0,1] is provided by:

  $$x_i = F^{-1}(u_i)$$

Exercise 1

Let $\underline{x}$ and $\underline{y}$ represent the outcomes of each of two dice. What is the probability of the following cases?

- $\{\underline{x} < \underline{y}\}$
- $\{\underline{x} < y\}$
- $\{x < \underline{y}\}$
- $\{x < y\}$

Verify the results by Monte Carlo simulations.

Exercise 2

- Assume that in a certain place on earth (specifically in the United Kingdom) and a certain period of the year a dry and a wet day are equiprobable and that in the different days the states (wet/dry) are independent. What is the probability that two consecutive days are wet under the following conditions?
  - ❑ Unconditionally.
  - ❑ If we know that the first day is wet.
  - ❑ If we know that the second day is wet.
  - ❑ If we know that one of the two days is wet.
  - ❑ If we know that one of the two days is dry.
- Verify the results by Monte Carlo simulations.
- Plot the distribution function of one day's state (wet/dry) (after introducing an appropriate random variable).
- Assuming that in a wet day the probability density function of the rainfall depth $\underline{x}$ (expressed in mm) is $f(x|\text{wet}) = \mathrm{e}^{-x}$, plot the probability distribution function $F(x)$.

Exercise 3

- Three engineers A, B and C are biding for a 1 000 000 € project and the evaluation committee, in order to make the fairest possible selection, decided to throw a die, instead of evaluating the proposal, the experience of engineers, etc.. If the outcome is 1 or 2 the projects goes to A, if it is 3 or 4, then B wins and if it is 5 or 6, then C wins. The dice is cast, but the announcement of the winner is going to be done the next day by the minister.
- Engineer A approaches the chairman of the committee and offers him 1000 € to accept his following request: "I know you are not allowed to tell me who wins; however, two of the three will lose. Therefore, B or C or both will lose. Please tell me just one of these two will lose". The committee member accepts and says that C will lose. Then engineer A offers another 1000 € to swap him with B.
- Prove that the strategy of engineer A is consistent with awareness of probability.
- Compare this strategy with another one, in which engineer A offers the same amount to convince the chairman to re-decide on A and B by tossing a coin.
- Verify your result with Monte Carlo simulation.

Note: A different utterance of this problem is known as the "three prisoners problem" (http://en.wikipedia.org/wiki/Three_Prisoners_problem), which has puzzled many. For example, Ben-Naim, 2008, devotes several pages in his book about entropy (including a whole appendix) to solve this problem. However, its solution can be done in two lines.

Expectation

- For a discrete random variable $\underline{x}$, taking on the values x_1, x_2, ..., x_w (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, if $g(\underline{x})$ is an arbitrary function of $\underline{x}$ (so that $g(\underline{x})$ is a random variable per se), we define the *expectation* or *expected value* or *mean* of $g(\underline{x})$ as

$$\mathrm{E}\left[g\left(\underline{x}\right)\right] \coloneqq \sum_{j=1}^{w} g\left(x_j\right) P\left(x_j\right)$$

- Likewise, for a continuous random variable $\underline{x}$ with density $f(x)$, the expectation is

$$\mathrm{E}\left[g\left(\underline{x}\right)\right] \coloneqq \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x$$

- For certain types of functions $g(\underline{x})$ we get very commonly used statistical parameters, as specified below:
  1. For $g(\underline{x}) = \underline{x}^r$, where $r = 0, 1, 2, \ldots$, the quantity $\mu'_r := \mathrm{E}[\underline{x}^r]$ is called the r*th moment* (or the r*th moment about the origin*) of $\underline{x}$. For $r = 0$, obviously the moment is 1.
  2. For $g(\underline{x}) = \underline{x}$, the quantity $\mu := \mu'_1 = \mathrm{E}[\underline{x}]$ (that is, the first moment) is called the *mean* of $\underline{x}$.
  3. For $g(\underline{x}) = (\underline{x} - \mu)^r$ where $r = 0, 1, 2, \ldots$, the quantity $\mu_r := \mathrm{E}[(\underline{x} - \mu)^r]$ is called the r*th central moment* of $\underline{x}$. For $r = 0$ and 1 the central moments are respectively 1 and 0. For
  4. For $g(\underline{x}) = (\underline{x} - \mu)^2$ the quantity $\gamma \equiv \sigma^2 := \mu_2 = \mathrm{E}[(\underline{x} - \mu)^2]$ is called the *variance* of $\underline{x}$ (also denoted as var[$\underline{x}$]); its square root σ (also denoted as std[$\underline{x}$] is called the standard deviation of $\underline{x}$.

Entropy

- For a **discrete random variable** $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_w$ (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, the *entropy* is defined as the expectation of the minus logarithm of probability (Shannon, 1948), i.e.:

$$\Phi[\underline{x}] := \mathrm{E}[-\ln P(\underline{x})] = -\sum_{j=1}^{w} P_j \ln P_j$$

- Extension of the above definition for the case of a **continuous random variable** $\underline{x}$ with probability density function $f(x)$, is possible, although not contained in Shannon's (1948) original work. This extension involves a (so-called) '*background measure*' with density $h(x)$, which can be any probability density, proper (with integral equal to 1) or improper (meaning that its integral does not converge); typically it is an (improper) Lebesgue density, i.e. a constant with dimensions $[h(x)] = [f(x)] = [x^{-1}]$, so that the argument of the logarithm function that follows be dimensionless. Thus, the entropy of a continuous variable $\underline{x}$ is (see e.g. Jaynes, 2003, p. 375):

$$\Phi[\underline{x}] := \mathrm{E}\left[-\ln \frac{f(\underline{x})}{h(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln \frac{f(x)}{h(x)} f(x) \mathrm{d}x$$

- It is easily seen that for both discrete and continuous variables the entropy $\Phi[\underline{x}]$ is a *dimensionless* quantity.
- The importance of the entropy concepts relies in the **principle of maximum entropy** (Jaynes, 1957); it postulates that the entropy of a random variable $\underline{x}$ should be at maximum, under some conditions, formulated as constraints, which incorporate the information that is given about this variable.
- This principle can be used for **logical inference** as well as for **modelling physical systems**; for example, the tendency of entropy to become maximal (Second Law of thermodynamics) can result from this principle.

Exercise 4

- Find the mean, variance and entropy of the variable $\underline{x}$ representing the outcome of a fair die. Show that the entropy of a fair die is greater than in any loaded die.
- Find the mean, variance and entropy of a variable $\underline{x}$ with uniform distribution in [0,1]. Show that this entropy is the maximum possible among all distributions in [0,1].
- Find the mean, variance and entropy of a variable $\underline{x}$ with exponential distribution. Show that this entropy is the maximum possible among all distributions in $[0,\infty)$ which have specified mean.
- Find the mean, variance and entropy of a variable $\underline{x}$ with normal distribution. Show that this entropy is the maximum possible among all distributions in $(-\infty,\infty)$ which have specified mean and variance.

Two variables: joint distribution and joint moments

- Here we provide definitions referring to a pair of two random variables $(\underline{x}, \underline{y})$.
- **Joint probability distribution function**: $F_{xy}(x, y) := P\{\underline{x} \le x, \underline{y} \le y\}$
- **Joint probability density function** : $f_{xy(x,y)} \coloneqq \frac{\partial^2 F_{xy}(x,y)}{\partial x\, \partial y}$
- **Marginal probability distribution functions** : $F_x(x) := P\{\underline{x} \le x\}$, $F_y(y) := P\{\underline{y} \le y\}$
- **Joint raw moment** of order $p + q$: $\mu'_{pq} \coloneqq \mathrm{E}[\underline{x}^p\ \underline{y}^q] = \int_{-\infty}^{\infty} x^p y^q\ f_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y$
- **Marginal first moments** (means): $\mu_x \coloneqq \mu'_{10}, \mu_y \coloneqq \mu'_{01}$
- **Joint raw moment** of order $p + q$:

$$\mu_{pq} \coloneqq \mathrm{E}\left[\left(\underline{x} - \mu_x\right)^p \left(\underline{y} - \mu_y\right)^q\right] = \int_{-\infty}^{\infty} \left(x - \mu_x\right)^p \left(y - \mu_y\right)^q f_{xy}(x, y) \mathrm{d}x\, \mathrm{d}y$$

- **Variances**: $\mathrm{var}\left[\underline{x}\right] \coloneqq \mathrm{E}\left[\left(\underline{x} - \mu_x\right)^2\right] \equiv \mu_{20} \equiv \gamma_x \equiv \sigma_x^2$; $\mathrm{var}\left[\underline{y}\right] \coloneqq \mu_{02} \equiv \gamma_y \equiv \sigma_y^2$
- **Covariance**: $\mathrm{cov}\left[\underline{x}, \underline{y}\right] \coloneqq \mathrm{E}\left[\left(\underline{x} - \mu_x\right)\left(\underline{y} - \mu_y\right)\right] \equiv \mu_{11} \equiv \sigma_{xy} = \mathrm{E}\left[\underline{x}\ \underline{y}\right] - \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right]$
- **Correlation coefficient**: $r_{xy} \coloneqq \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- **Independent variables**: $F_{xy}(x, y) = F_x(x)\ F_y(y)$; $f_{xy}(x, y) = f_x(x)\ f_y(y)$
- **Uncorrelated variables**: $\sigma_{xy} = 0$, $r_{xy} = 0$, $\mathrm{E}[\underline{x}\ \underline{y}] = \mathrm{E}[\underline{x}]\ \mathrm{E}[\underline{y}]$

Correlation and climacogram

- **Linear combinations of random variables**: $\mathrm{E}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1\mathrm{E}\left[\underline{x}_1\right]+a_2\mathrm{E}\left[\underline{x}_2\right]$, $\mathrm{var}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1^2\mathrm{var}\left[\underline{x}_1\right] + a_2^2\mathrm{var}\left[\underline{x}_2\right] + 2a_1a_2\mathrm{cov}\left[\underline{x}_1,\underline{x}_2\right]$
- It follows that: $\mathrm{Var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}+\frac{\underline{x}_2}{\sigma_2}\right)\right] = \frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1-\mu_1}{\sigma_1}+\frac{\underline{x}_2-\mu_2}{\sigma_2}\right)^2\right] = \frac{1}{2}+\frac{1}{2}\mathrm{Cov}\left[\frac{\underline{x}_1}{\sigma_1},\frac{\underline{x}_2}{\sigma_2}\right]$
- Likewise: $\mathrm{Var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}-\frac{\underline{x}_2}{\sigma_2}\right)\right] = \frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1-\mu_1}{\sigma_1}-\frac{\underline{x}_2-\mu_2}{\sigma_2}\right)^2\right] = \frac{1}{2}-\frac{1}{2}\mathrm{Cov}\left[\frac{\underline{x}_1}{\sigma_1},\frac{\underline{x}_2}{\sigma_2}\right]$
- Thus, $r_{12} = \frac{\mathrm{Cov}\left[\underline{x}_1,\underline{x}_2\right]}{\sigma_1\sigma_2} = \mathrm{cov}\left[\frac{\underline{x}_1}{\sigma_1},\frac{\underline{x}_2}{\sigma_2}\right] = 2\ \mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}+\frac{\underline{x}_2}{\sigma_2}\right)\right] - 1 = 1 - 2\ \mathrm{var}\left[\frac{\underline{x}_1}{\sigma_1}-\frac{\underline{x}_2}{\sigma_2}\right]$
- As the variance is by definition non-negative, it follows that $-1 \le r_{12} \le 1$; the value r_{12} =0 corresponds to uncorrelated variables, while positive or negative r_{12} corresponds to positively or negatively correlated variables, respectively.
- The same information as in r_{12} is provided by the quantity $\rho_{12} \coloneqq \mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1}+\frac{\underline{x}_2}{\sigma_2}\right)\right]$, for which it is easily seen that $0 \le \rho_{12} \le 1$; the value ρ_{12} =1/2 corresponds to uncorrelated variables, while values of ρ_{12} greater or less than ½ correspond to positively or negatively correlated variables, respectively.
- The notion of ρ_{12} could be readily expanded to many variables. Assuming that all variables are identically distributed and multiplying by the common variance σ^2, we define the so-called **climacogram**, $\gamma_\kappa \coloneqq \mathrm{var}\left[\underline{X}_k/k\right]$, where $\underline{X}_k \coloneqq \underline{x}_1 + \cdots + \underline{x}_\kappa$ and $0 \le \gamma_\kappa \le \sigma^2$.

Many variables and stochastic processes

- A **stochastic process** is a family of infinitely many random variables indexed by a (regular) variable, which takes values from an index set T, typically representing time. We distinguish between:
  - A **continuous-time stochastic process** $\underline{x}(t)$, when time is continuous, e.g. $T = [0, \infty)$.
  - A **discrete-time stochastic process** $\underline{x}_i$, when time is discrete, e.g., $T = \{0, 1, 2, \ldots\}$.
- **Time series or sample function**: a realization, x_i, of a stochastic process, $\underline{x}_i$ or $\underline{x}(t)$, at a finite set of discrete time instances i (or t_i). (**Caution**: A stochastic process is a family of random variables, infinitely many for discrete time processes and uncountably infinitely many for continuous time processes. On the other hand, a time series is a finite sequence of numbers).
- **First order distribution function** of the process: $F(x; t) := P\{\underline{x}(t) \le x\}$
- **Second order distribution function** : $F(x_1, x_2; t_1, t_2) := P\{\underline{x}(t_1) \le x_1, \underline{x}(t_2) \le x_2\}$
- ***n*th order distribution function**: $F(x_1, \ldots, x_n; t_1, \ldots, t_n) := P\{\underline{x}(t_1) \le x_1, \ldots, \underline{x}(t_n) \le x_n\}$
- **Mean**: $\mu(t) := \mathrm{E}[\underline{x}(t)]$
- **Autocovariance**: $c(t; h) := \mathrm{Cov}[\underline{x}(t), \underline{x}(t+h)] = \mathrm{E}[(\underline{x}(t) - \mu(t))\,(\underline{x}(t+h) - \mu(t+h))]$
- **Cross-covariance of two processes** $\underline{x}(t)$ and $\underline{y}(t)$: $c_{xy}(t; h) := \mathrm{cov}[\underline{x}(t), \underline{y}(t+h)]$

Stationarity

- Central to the notion of a stochastic process are the concepts of *stationarity* and *nonstationarity*, two widely misunderstood and misused concepts (see Koutsoyiannis and Montanari, 2014), whose definitions apply only to stochastic processes (thus, e.g., a time series cannot be stationary, nor nonstationary).
- A process is called **(strict-sense**) **stationary** if its statistical properties are invariant to a shift of time origin, i.e. the processes $\underline{x}(t)$ and $\underline{x}(t')$ have the same statistics for any t and t' (see further details in Papoulis, 1991; see also further explanations in Koutsoyiannis, 2006, 2011 and Koutsoyiannis and Montanari, 2015). Conversely, a process is nonstationary if some of its statistics are changing through time and their change is described as a deterministic function of time.
- A stochastic process is called **wide-sense stationary** if its mean is constant and its autocovariance depends on time difference only, i.e.

$$\mathrm{E}[\underline{x}(t)] = \mu = \text{constant}, \quad \mathrm{E}[(\underline{x}(t) - \mu)\ (\underline{x}(t + \tau) - \mu)] = c(\tau)$$

- Convenient tools for a stationary process, which can replace auto- and cross-covariance, are the following:
    - **Climacogram**: $\gamma(k) := \mathrm{var}[\underline{X}(k)/k]$, where $\underline{X}(k) \coloneqq \int_0^k \underline{x}(t)\mathrm{d}t$.
    - **Cross-climacogram** of two stationary processes $\underline{x}(t)$ and $\underline{y}(t)$:
    $\gamma_{xy}^{\eta}(k) \coloneqq \sigma_x \sigma_y \ \mathrm{var}\left[\frac{\underline{X}(k)}{k\sigma_x} + \frac{\underline{Y}((\eta+1)k) - \underline{Y}(\eta k)}{k\sigma_y}\right]$, where $\underline{Y}(k) \coloneqq \int_0^k \underline{y}(t)\mathrm{d}t$ and η is lag.

Ergodicity

- Stationarity is also related to *ergodicity*, which in turn is a prerequisite to make inference from data, that is, induction. Without ergodicity inference from data would not be possible. Ironically, several studies use time series data to estimate statistical properties, as if the process were ergodic, while at the same time what they (cursorily) estimate may falsify the ergodicity hypothesis (see example on p. 22).
- While ergodicity is originally defined in dynamical systems (e.g. Mackey, 1992, p. 48), the ergodic theorem (e.g. Mackey, 1992 p. 54) allows redefining ergodicity within the stochastic processes domain (Papoulis 1991 p. 427; Koutsoyiannis 2010) in the following manner: A stochastic process $\underline{x}(t)$ is ergodic if the time average of any (integrable) function $g(\underline{x}(t))$, as time tends to infinity, equals the true (ensemble) expectation $\mathrm{E}[g(\underline{x}(t))]$, i.e., $\lim_{T\to\infty} \frac{1}{T}\int_0^T g\left(\underline{x}(t)\right) dt = \mathrm{E}[g(\underline{x}(t))]$.
- If the system that is modelled in a stochastic framework has deterministic dynamics (meaning that a system input will give a single system response, as happens for example in most hydrological models) then a theorem applies (Mackey 1992, p. 52), according to which a dynamical system has a stationary probability density *if and only if* it is ergodic. Therefore, a stationary system is also ergodic and vice versa, and a nonstationary system is also non-ergodic and vice versa.
- If the system dynamics is stochastic (a single input could result in multiple outputs), then ergodicity and stationarity do not necessarily coincide. However, recalling that a stochastic process is a model and not part of the real world, we can always conveniently device a stochastic process that is ergodic (see example in Koutsoyiannis and Montanari, 2015).
- In conclusion, from a practical point of view ergodicity can always be assumed when there is stationarity.

A note on statistical estimation

- Models are human inventions and not part of the real world. They are characterized by their mathematical structure and their parameters. The field of stochastics allows both testing the model structure and estimating the parameters, based on observation data. This is induction in practice and it is made possible by virtue of the ergodic theorem.
- We should be aware of the differences between **three concepts** related to a single parameter θ:
  - The **true** but unknown **value** θ (often called “population” parameter) .
  - The **estimator** $\underline{\hat{\theta}}$, which is a random variable depending on the stochastic process of interest $\underline{x}(t)$. $\underline{\hat{\theta}}$ is a model per se, not a number.
  - The **estimate** $\hat{\theta}$ which is a number calculated by using the observations and the estimator.
- Characteristic statistics of the estimator $\underline{\hat{\theta}}$ are its **bias**, $\mathrm{E}\left[\underline{\hat{\theta}}\right] - \theta$, and its **variance** $\mathrm{var}\left[\underline{\hat{\theta}}\right]$. When $\mathrm{E}\left[\underline{\hat{\theta}}\right] = \theta$ the estimator is called unbiased.
- As an example, the standard estimator of the **mean** from a finite set of random variables $\underline{x}_i$ (sample of size n), taken from a stochastic process $\underline{x}(t)$ at discrete time instances *i*, is $\underline{\hat{\mu}} := \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$; it is easy to show that it is **unbiased**.
- However, the the standard estimator of the **variance** from the same set of random variables $\underline{x}_i$ is $\underline{\hat{\gamma}} := \frac{1}{n-1} \sum_{i=1}^{n} \left(\underline{x}_i - \underline{\hat{\mu}}\right)^2$; even though it is often called unbiased, it is **biased**, unless $\underline{x}_i$ are independent, which is rarely the case in geophysics (see Koutsoyiannis, 2016).

References

- Ben-Naim, A., *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific Pub., Singapore, 384 pp., 2008.
- Berry, M., Regular and irregular motion, in *Topics in nonlinear dynamics: A tribute to Sir Edward Bullard*, edited by S. Jorna, American Institute of Physics, New York, 1978 (pp. 16-120)
- Gauch, H.G., Jr., *Scientific Method in Practice*, Cambridge University Press, Cambridge, 2003.
- Jaynes, E.T. Information theory and statistical mechanics, *Physical Review*, 106 (4), 620-630, 1957.
- Jaynes, E.T. *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, 728 pp., 2003.
- Kolmogorov, A. N., Grundbegrijfe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Math.* (2), Berlin, 1933; 2nd English Edition: Foundations of the Theory of Probability, 84 pp. Chelsea Publishing Company, New York, 1956.
- Koutsoyiannis, D., Nonstationarity versus scaling in hydrology, *Journal of Hydrology*, 324, 239–254, 2006.
- Koutsoyiannis, D., A random walk on water, *Hydrology and Earth System Sciences*, 14, 585–601, 2010.
- Koutsoyiannis, D., Hurst-Kolmogorov dynamics and uncertainty, *Journal of the American Water Resources Association*, 47 (3), 481–495, 2011.
- Koutsoyiannis, D., *Probability and statistics for geophysical processes*, National Technical University of Athens, Athens, 2008 (itia.ntua.gr/1322/).
- Koutsoyiannis, D., Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrological Sciences Journal*, 61 (2), 225–244, doi: 10.1080/02626667.2015.1016950, 2016.
- Koutsoyiannis, D.. and Montanari, A., Negligent killing of scientific concepts: the stationarity case, *Hydrological Sciences Journal*, 60 (7-8), 1174–1183, doi:10.1080/02626667.2014.959959, 2015.
- Mackey, M.C., *Time's Arrow: The Origins of Thermodynamic Behavior*, Dover, Mineola, NY, USA, 175 pp., 2003.
- Papoulis, A., *Probability and Statistics*, Prentice-Hall, New Jersey, 1990.
- Ripley, B. D., *Stochastic Simulation*, Wiley, New York, 1987.
- Ruelle, D., Microscopic fluctuations and turbulence, *Phys. Letters*, 72A, 81-82, 1979.
- Ruelle, D., Chance and chaos, Princeton University Press, 1991.
- Shannon, C.E. The mathematical theory of communication, *Bell System Technical Journal*, 27 (3), 379-423, 1948.