Stochastic Systems — Discrete Time Systems

Simone Santini

Class notes, Academic Year 2021/2022

1 Random variables

A **probability space** is a triple (Ω, σ, μ) where

probability space

i) Ω is a set of *outcomes* (we shall use the notation $\omega \in \Omega$ for its elements).

ii) σ is a σ-algebra on Ω, that is a non-empty set of subsets of Ω ($\sigma \subseteq 2^\Omega$) such that

- **a)** $\Omega \in \sigma$
- **b)** $A \in \sigma \implies \Omega \backslash A \in \sigma$ (we shall use the notation A^c for $\Omega \backslash A$)
- **c)** If $A_n \in \sigma$ for all $n \in \mathbb{N}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \sigma$.

iii) $\mu : \sigma \to [0, 1]$ is such that

- **a)** $\mu(\Omega) = 1$
- **d)** Given A_n, $n \in \mathbb{N}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, then

$$\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n=0}^{\infty} \mu(A_n) \tag{1}$$

Condition ii) implies that $\bigcap_{n\in\mathbb{N}} A_n \in \sigma$, since

$$\bigcap_{n\in\mathbb{N}} A_n = \Big(\bigcap_{n\in\mathbb{N}} A_n^c\Big)^c \tag{2}$$

A subset $A \subset \Omega$ with $A \in \sigma$ is called an **event**. It is possible to show (we shall not do it here) that the sum in (1) is well defined, that it, it is independent on the way we index the sets A_n. With these properties, it is possible to take limit:

event

Theorem 1.1. *Is A_n, $n \in \mathbb{N}$ is a sequence of subsets of Ω and $\lim_{n\to\infty} A_n = A$, then $\lim_{n\to\infty} \mu(A_n) = \mu(A)$.*

Let M be a countable set. A (discrete) *random variable* is a function $X : \Omega \to M$ such that, for all $m \in M$,

random variable

$$\Big\{X = m\Big\} \triangleq \Big\{\omega \in \Omega \Big| X(\omega) = m\Big\} \in \sigma \tag{3}$$

That is, a random variable is a mathematical objet that associates values of M to outcomes in such a way that the back-image of any value of M is an event. The probability of the event $\{X = m\}$ is

$$P_X(m) \triangleq \mu\big(\{X = m\}\big) \tag{4}$$

The properties of μ induce corresponding properties in P_X:

i) $m \in M \implies P_X(m) \geq 0$

ii) $\sum_{m\in M} P_x(m) = 1$

If Ω and M are uncountable, the general principles are the same, but the definitions and the conditions are technically more complex, and we shall not go into the details here. In this case, P_X is a **probability density function** (henceforth: PDF), events are subsets of Ω of non-zero measure, and probabilities are integrals of P_X over sets of finite measure determined by images of events. In this case, the normalization condition is

probability density

$$\int_\Omega P_X(x)dx = 1 \tag{5}$$

If $\Omega = \mathbb{R}$ (as we shall often assume) we have

$$\int_{-\infty}^{\infty} P_X(x)dx = 1 \tag{6}$$

For continuous variables on $\mathbb{R}$ one can define the **cumulative probability function**, that is, the probability that X be at most x:

cumulative probability

$$\mathcal{P}(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} P_X(u)du \tag{7}$$

Note that

$$P_X(x) = \frac{\partial}{\partial x}\mathcal{P}(x). \tag{8}$$

From this and the positivity condition one can show that $\mathcal{P}$ is monotonically non-decreasing and that

$$\lim_{x\to-\infty} \mathcal{P}(x) = 0 \qquad \lim_{x\to\infty} \mathcal{P}(x) = 1 \tag{9}$$

A whole function such as P_X is a cumbersome object, and difficult to work with; it is easier to work with an enumerable set of numbers that characterizes the function completely. **Statistical moments** are such quantities. The moment of order n of the variable X is defined as

statistical moments

$$\langle X^n \rangle = \int_M x^n P_X(x) dx \tag{10}$$

In general, given a function f defined on M, we define

$$\langle f(X) \rangle = \int_M f(x) P_X(x) dx \tag{11}$$

The nth moment is obtained for $f(x) = x^n$.

The first order moment $\langle X \rangle$ is called the **average**, or the **expected value** of X, while

expected value

$$\sigma^2 = \langle X^2 \rangle - \langle X \rangle^2 \tag{12}$$

is its **variance**; the square root of the variance, σ, is the **standard deviation** of X.

variance
standard deviation

Not all distributions have finite moments, that is, the integral (10) may fail to converge. If the moments are finite, then they completely characterize the PDF. To show this, we introduce the **characteristic function** $\tilde{P}_X(\omega)$[1] of a PDF P_X:

characteristic function

$$\tilde{P}_X(\omega) = \langle e^{i\omega x} \rangle = \int_M e^{i\omega x} P_X(x)\, dx \tag{13}$$

This is simply the Fourier transform of P_X, so the PDF can be recovered from its characteristic function as

$$P_X(x) = \frac{1}{2\pi} \int e^{-i\omega x} \tilde{P}_X(\omega)\, d\omega \tag{14}$$

The relation with the moments becomes evident by taking the Taylor expansion of the exponential:

$$e^{i\omega x} = \sum_n^\infty \frac{(i\omega x)^n}{n!} \tag{15}$$

[1] There is a possible confusion here: we have already used the symbol ω to indicate an ocurrence ($\omega \in \Omega$), and now we are using it to indicate the pulsation variable in the Fourier Transform. Unfortunately, both usages are standard and very common (there are many more mathematical concepts than letters available in the Latin and Greek alphabets, and moreexotic symbols are hard to come by in LaTeX). Fortunately, ew shall not need the set Ω anymore, so from now on ω will indicate the Fourier transform variable.

Introducing this into (13) we get

$$\tilde{P}_X(\omega) = \sum_n \frac{(i\omega)^n}{n!} \int x^n P_X(x)\, dx = \sum_n \frac{(i\omega)^n}{n!} \langle X^n \rangle \tag{16}$$

A useful consequence of this expansion is that the moments of P_X can be obtained by differentiating $\tilde{P}_X$:

$$\langle X^n \rangle = \lim_{\omega \to 0} (-i)^n \frac{\partial^n}{\partial \omega^n} \tilde{P}_X(\omega) \tag{17}$$

* * *

joint probability

conditional probability
independence

The **joint probability** of two random variables X_1 and X_2, indicated as $P_{X_1 \cap X_2}(x_1, x_2)$ measures the simultaneous probability that X_1 and X_2 take the values x_1 and x_2, respectively. The **conditional probability** $P_{X_1|X_2}(x_1|x_2)$ denotes the probability that X_1 take value x_1 contitioned to the fact that X_2 takes value x_2. Two variables are **independent** if for all x_1, x_2 $P_{X_1|X_2}(x_1 x_2) = P_{X_1}(x_1)$, that is, if knowing the value of X_2 does not change the distribution of X_1. Joint and conditional probabilities are related through Bayes's theorem:

$$P_{X_1 \cap X_2}(x_1, x_2) = P_{X_1|X_2}(x_1|x_2) P_{X_2}(x_2) = P_{X_2|X_2}(x_2|x_1) P_{X_1}(x_1) \tag{18}$$

Similar definitions can be given for probability density functions.

1.1 Useful Probability Distributions

Gaussian distribution

A variable X follows a **Gaussian** (or *normal*) distribution if

$$P_X(x) = N_X(x) = N(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{\sigma^2}\Big) \tag{19}$$

(Figure 1) or, equivalently, it has characteristic function

$$\tilde{P}_X(\omega) = \tilde{N}(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} P_X(x)\, dx = \exp\Big(i\omega\mu - \frac{\omega^2\sigma^2}{2}\Big) \tag{20}$$

The mean of the distribution is $\langle X \rangle = \mu$, and its variance is σ^2. Note that, for $\mu = 0$, the

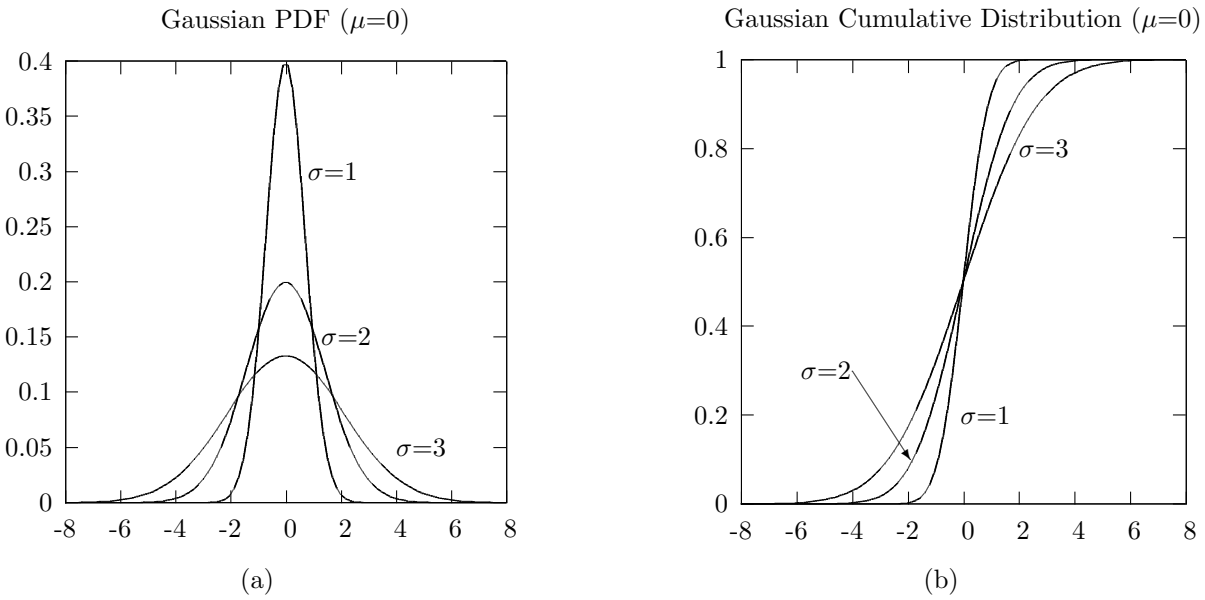

Figure 1: The Gaussian PDF (a) and the corresponding cumulative distribution (b) for various values of σ; in all cases it is $\mu = 0$.

characteristic function has also the functional form of a Gaussian, a fact that will come handy in the following. In this special case ($\mu = 0$) the moments are given by

$$\langle X^n \rangle = \int_{-\infty}^{\infty} x^n \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{\sigma^2}\Big)\, dx = \begin{cases} \frac{2^{\frac{n}{2}}\sigma^n}{\sqrt{\pi}}\Gamma\Big(\frac{n+1}{2}\Big) & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \tag{21}$$

where Γ is Euler's Gamma function. An important moment is

$$\langle X^2 \rangle = \sigma^2 + \langle X \rangle. \tag{22}$$

One important property of the Gaussian distribution, vis-à-vis the Central Limit Theorem (which we shall consider in the following), is that it is **stable**: if X, Y are Gaussians, and $a, b \in \mathbb{R}$, then $aX + bY$ is also Gaussian.

stable distribution

Let X and Y be two Gaussian-distributed variables with zero mean and variance σ_1^2 and σ_2^2, respectively. Then

$$\begin{aligned} \tilde{N}_X(\omega) &= \exp\big(-\frac{\omega^2\sigma_1^2}{2}\big) \\ \tilde{N}_Y(\omega) &= \exp\big(-\frac{\omega^2\sigma_2^2}{2}\big) \end{aligned} \tag{23}$$

and consequently

$$\tilde{N}_X(\omega)\tilde{P}_Y(\omega) = \exp\big(-\frac{\omega^2(\sigma_1^2+\sigma_2^2)}{2}\big) \tag{24}$$

that is, the product of the characteristic function of two Gaussian distributions is still the characteristic function of a Gaussian distribution. From the general properties of the Fourier transform, we deduce that the convolution of two Gaussians distributions is still a Gaussian distribution.

* * *

The Gaussian distribution is defined for all $x \in \mathbb{R}$, but in the model of many interesting phenomena the variable assumes only positive values in such a way that the probability that $x = 0$ is 0 and, after reaching a maximum, decreases rapidly for high values of x. Examples of this kind of phenomena abound and are of the most diverse nature, from the length of messages in internet fori, to the price of hotels or the size of the fragments resulting from a collision. In these cases, negative values are out of the question, so we can't model them using a Gaussian distribution for which $N_X(x) > 0$ for all $x \in \mathbb{R}$.

All these phenomena can be modeled by means of a **logonormal distribution**. A variable X

logonormal distribution

has logonormal distribution if $\log X$ has normal (viz. Gaussian) distribution. Let Φ and ϕ be the cumulative distribution and the density of a normally distributed variable with 0 mean and unit variance ($\mathcal{N}(0,1)$), and assume $\log X \sim \mathcal{N}(\mu,\sigma)$, i.e. $\log X$ has a normal distribution with mean μ and variance σ^2. Then

$$\begin{aligned} P_X(x) &= \frac{d}{dx}\mathcal{P}_X(x) = \frac{d}{dx}\mathbb{P}[X \le x] \\ &= \frac{d}{dx}\mathbb{P}[\log X \le \log x] \\ &= \frac{d}{dx}\Phi\big[\frac{\log x - \mu}{\sigma}\big] \\ &= \phi\big[\frac{\log x - \mu}{\sigma}\big]\frac{d}{dx}\big[\frac{\log x - \mu}{\sigma}\big] \\ &= \frac{1}{\sigma x}\phi\big[\frac{\log x - \mu}{\sigma}\big] \\ &= \frac{1}{\sqrt{2\pi}\sigma x}\exp\Big[-\frac{(\log X - \mu)^2}{2\sigma^2}\Big] \end{aligned} \tag{25}$$

Figure 2 shows the behavior of the logonormal PDF for various values of σ and $\mu = 0$ Note that μ

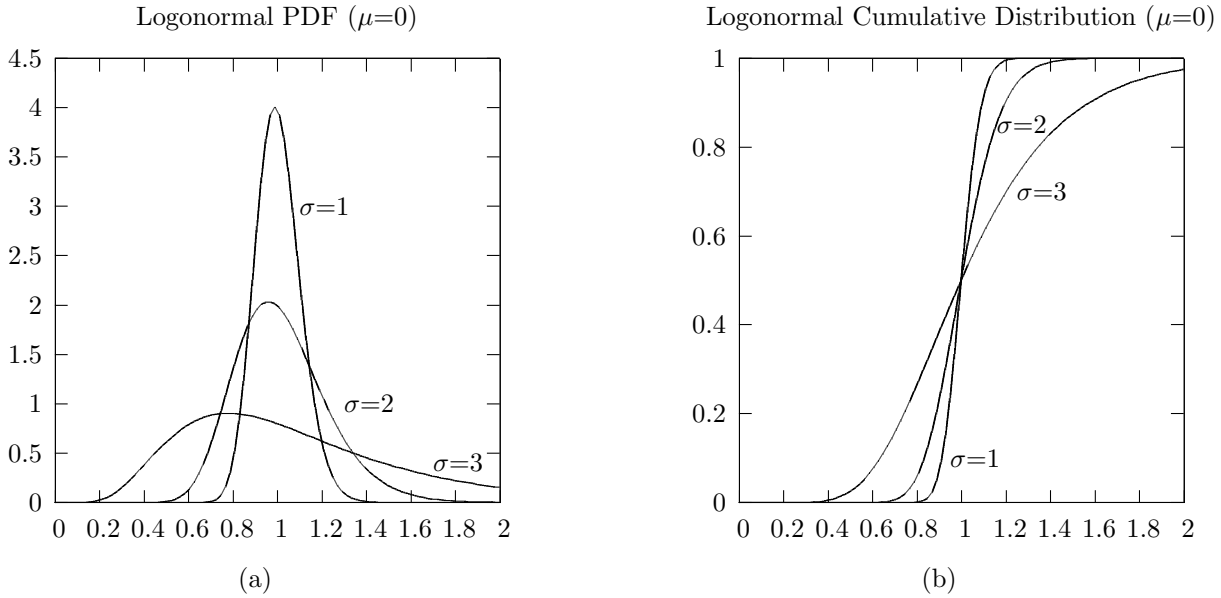

Figure 2: The Logonormal PDF (a) and the corresponding cumulative distribution (b) for various values of σ; in all cases it is $\mu = 0$.

and σ are the mean and variance of $\log X$, *not* of X. To distinguish them, I shall indicate the mean and the variance of X as m and v, respectively.

The moments of X are given by

$$\langle X^n \rangle = \int_0^\infty x^n P_X(x) dx = \exp\big(n\mu + \frac{n^2\sigma^2}{2}\big) \tag{26}$$

as can be verified by replacing $z = \frac{1}{\sigma}\big[\log X - (\mu + n\sigma^2)\big]$ in the integral. From this we have

$$\begin{aligned} m = \langle X \rangle &= \exp\big(\mu + \frac{\sigma^2}{2}\big) \\ \langle X^2 \rangle &= \exp\big(2\mu + 2\sigma^2\big) \\ v = \langle X^2 \rangle - \langle X \rangle^2 &= \exp(2\mu + \sigma^2)(e^{\sigma^2} - 1) \end{aligned} \tag{27}$$

From these equality, one can derive the values of μ and σ^2 for desired m and v:

$$\mu = \log \frac{m}{\sqrt{1 + \frac{v}{m^2}}} \quad \sigma^2 = \log\Big(1 + \frac{v}{m^2}\Big) \tag{28}$$

The characteristic function $\langle \exp(i\omega x) \rangle$ is defined, but if we try to extend it to complex variables, $\langle \exp(sx) \rangle$, $s \in \mathbb{C}$ is not defined for any s with a negative imaginary part. This entails that the characteristic function is not analytical in the origin and, consequently, it can't be represented as an infinite convergent series. In particular, the formal Taylor series

$$\sum_n \frac{(i\omega x)^n}{n!} \langle x^n \rangle = \sum_n \frac{(i\omega x)^n}{n!} \exp\big(n\mu + \frac{n^2\sigma^2}{2}\big) \tag{29}$$

diverges

* * *

Other positive variables follow a different distribution, one in which the value 0 is the most probable, and the probability decreases sharply as x increases, In these cases, the variable x can be modeled using an **exponential** distribution (Figure 3).

Exponential distribution

$$P_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{30}$$

If the variable can take negative values, then the distribution takes the name of **Laplace** distribution

Laplace distribution

$$P_X(x) = \frac{\lambda}{2} e^{-\lambda|x|} \tag{31}$$

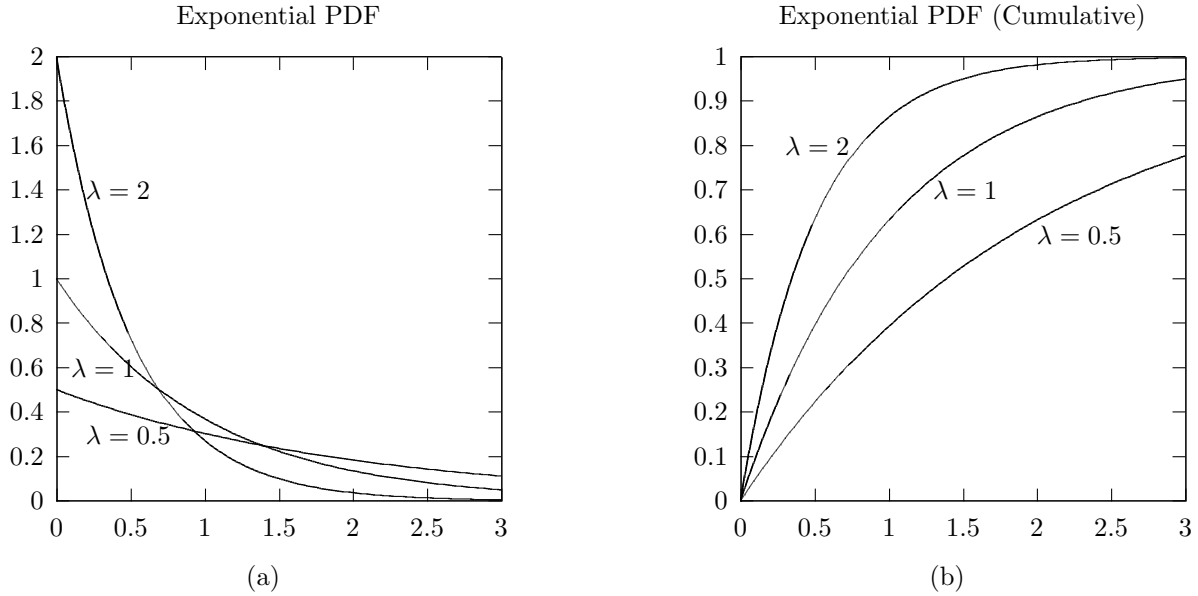

Figure 3: The exponential PDF (a), an the corrsponding cumulative distribution in (b) for various values of λ.

Its characteristic function is

$$\tilde{P}_X(\omega) = \frac{\lambda^2}{\lambda^2 + \omega^2} \tag{32}$$

and its moments

$$\langle X^n \rangle = \frac{1}{\lambda^n}\Gamma(n+1) \tag{33}$$

* * *

Uniform distribution

A **uniform** or *flat* distribution assigns the same probability density to each point in Ω. So, if $\Omega = [a, b]$,

$$P_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

The characteristic function of the uniform distribution is

$$\tilde{P}_X(\omega) = \frac{e^{i\omega b} - e^{i\omega a}}{i\omega(b-a)} \tag{35}$$

and its moments

$$\langle X^n \rangle = \frac{1}{n+1} \frac{b^{n+1} - a^{n+1}}{b-a} \tag{36}$$

* * *

Cauchy distribution

A **Cauchy**, or **Lorentz** distribution has PDF

$$P_X(x) = \frac{1}{\pi} \frac{\gamma}{x^2 + \gamma^2} \tag{37}$$

where γ is a positive parameter, and characteristic function

$$\tilde{P}_X(\omega) = e^{-\gamma|\omega|} \tag{38}$$

If one tries to compute the moments using the definition

$$\langle X^n \rangle = \frac{\gamma}{\pi} \int \frac{x^n}{x^2 + \gamma^2} \, dx \tag{39}$$

then, since the integrand behaves as x^{n-2} for $x \to \infty$, one observes that they diverge for $n \geq 1$. This limits the usefulness of this distribution as a model of real phenomena (which typically have finite moments), and in practice one "truncates" the distribution to a finite interval $[a, b]$.

* * *

Lèvy distribution

We have mentioned that one important property of the Gaussian distribution is the preservation of the functional form of their characteristic function under multiplication, as in (24). The Gaussian distribution is not the msot general distribution with this property (although it is the only one with this property *and* finite moments): it is shared by the family of **Lévy distributions**. Lévy distributions depend on four parameters: α (Lévy index), β (skewness), μ (shift), and σ (scale), and they are defined through their characteristic function:

$$\tilde{P}_{\alpha,\beta}(\omega; \mu, \sigma) = \int_{-\infty}^{\infty} e^{i\omega x} P_{\alpha,\beta}(x; \mu, \sigma) \, dx \triangleq \exp\left[i\mu\omega - \sigma^\alpha |\omega|^\alpha \left(1 - i\beta \frac{\omega}{|\omega|} \Phi \right) \right] \tag{40}$$

where

$$\Phi = \begin{cases} \tan \frac{\alpha\pi}{2} & \alpha \neq 1, 0 < \alpha < 2 \\ -\frac{2}{\pi} \ln |x| & \alpha = 1 \end{cases} \tag{41}$$

the four parameters determine the shape of the distribution. Of these, α and β play a major rôle in this note, while μ and σ can be eliminated through proper scale and shift transformations (much like mean and variance for the Gaussian distribution):

$$P_{\alpha,\beta}(x;\mu,\sigma) = \frac{1}{\sigma} P_{\alpha,\beta}(\frac{x-\mu}{\sigma};0,1) \tag{42}$$

From now on, I shall therefore ignore μ and σ and refer to the distribution as $P_{\alpha,\beta}(x)$. Note the symmetry relation

$$P_{\alpha,-\beta}(x) = P_{\alpha,\beta}(x) \tag{43}$$

The distributions with $\beta = 0$ are symmetric, and these are the ones that are the most relevant in this context. The closed form of $P_{\alpha,\beta}$ is known only for a few cases. If $\alpha = 2$ one obtains the Gaussian distribution (β is irrelevant, since $\Phi = 0$); if $\alpha = 1, \beta = 0$ one obtains the Cauchy distribution, and for $\alpha = 1/2, \beta = 1$, the Lévy-Smirnov distribution

$$P_{1/2,1}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} x^{-\frac{3}{2}} \exp\left(-\frac{1}{2x}\right) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{44}$$

The most important property in this context is the asymptotic behavior of $P_{\alpha,\beta}$ which is given by the power law

$$P_{\alpha,0}(x) \sim \frac{C(\alpha)}{|x|^{1+\alpha}} \tag{45}$$

with

$$C(\alpha) = \frac{1}{\pi} \sin\left(\frac{\pi\alpha}{2}\right) \Gamma(1+\alpha) \tag{46}$$

This power law behavior entails that arbitrarily large values are relatively probable (compared with the exponential decay of the Gaussian). Consequently, as can be expected, $\langle X^2 \rangle$ diverges for $\alpha < 2$.

* * *

Dirac δ distribution

The **Dirac delta distribution** is a pathological distribution useful in many contexts; for example, when dealing with certainty in a probabilistic framework, or when analyzing discrete random variables in a context created for continuous ones. It is the distribution of a continuous variable that takes one specific value with probability one. The distribution is indicated as

$$P_X(x) = \delta(x - x_0) \tag{47}$$

where $\delta(\cdot)$ is the "Dirac delta". The characteristic function of the distribution is

$$\tilde{P}_X(\omega) = \exp(i\omega x_0). \tag{48}$$

The function $\delta(x)$ is zero everywhere except for $x = 0$, and

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1 \tag{49}$$

This property entails $\delta(ax) = \delta(x)/a$. Also

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0) \tag{50}$$

from which we derive

$$\langle x^n \rangle = x_0^n \tag{51}$$

* * *

binomial distribution

Unlike the previous distribution, the **binomial distribution** is defined for discrete variables, in particular for a variable X that can take two values, the first one with probability p, and the second one with probability $1-p$. Suppose, for example, that we play a game in which, at each turn, I have a probability p of winning and $1 - p$ of losing (think of head-and-tails game with a tricked coin). If we play N rounds of the game, what is the probability that I win exactly n times? This turns out to be

$$P(X = n) = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n} = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \tag{52}$$

which is precisely the binomial distribution. Its characteristic function is

$$\tilde{P}(\omega) = (1 - p + pe^{i\omega})^N \tag{53}$$

from which the moments can be derived. For example

$$\langle X \rangle = \lim_{\omega \to 0} \frac{d\tilde{P}}{d\omega} = \lim_{\omega \to 0} pNe^{i\omega}(1 - p + pe^{i\omega})^{N-1} = pN \tag{54}$$

* * *

Poisson distribution

An important and common distribution, one that appears as a limiting case of many finite processes, is the **Poisson Distribution**. Its importance will probably be more evident if we derive it

as a limiting case in some examples.

Example I:
Consider events that may happen at any moment in time (the events are punctual: they have no duration). Divide the time-line in small intervals of duration Δt, so short that the probability that two or more events will take place in the same interval is negligible. Assume that the probability that *one* event take place in $[t, t+\Delta t)$ is constant, and proportional to the length of the interval:

$$P(1;\Delta t) = \lambda \Delta t \tag{55}$$

and, because no two events happen in the same interval,

$$P(0;\Delta t) = 1 - \lambda \Delta t \tag{56}$$

Let $P(0;t)$ be the probability that no event has taken place up to time t. Then

$$P(0;t+\Delta t) = P(0;t)(1-\lambda \Delta t) \tag{57}$$

Rearranging the terms we get

$$\frac{P(0;t+\Delta t) - P(0;t)}{\Delta t} = -\lambda P(0;t) \tag{58}$$

and, taking the limit for $\Delta t \to 0$

$$\frac{\partial}{\partial t} P(0;t) = -\lambda P(0;t) \tag{59}$$

that is, $P(0;t) = C\exp(-\lambda t)$ or, considering the boundary condition $P(0,0) = 1$ (nothing happens in no time),

$$P(0;t) = e^{-\lambda t} \tag{60}$$

This takes care of the case in which no event takes place before time t. On to the general case: what is the probability that (exactly) n events take place by time t?. We have n events by time $t+\Delta t$ if either (1) we had n events up to time t and no event ocurred in $[t, t+\Delta t]$, or (2) there were $n-1$ events at t and one event ocurred in $[t, t+\Delta t]$. This leads to

$$P(n;t+\Delta t) = (1-\lambda \Delta t)P(n;t) + \lambda \Delta t P(n-1;t) \tag{61}$$

rearranging and taking the limit $\Delta t \to 0$, we have

$$\frac{\partial}{\partial t} P(n;t) + \lambda P(n;t) = \lambda P(n-1;t) \tag{62}$$

In order to transform this equation into a more manageable form, we look for a function that, multiplied by the left-hand side, transforms it into the derivative of a product. That is, we look for a function $\mu(t)$ such that

$$\mu(t)\left[\frac{\partial P}{\partial t} + \lambda P\right] = \frac{\partial}{\partial t}\big[\mu(t)P\big] \tag{63}$$

It is easy to verify that $\mu(t) = \exp(\lambda t)$ fits the bill. Equation (62) therefore becomes

$$\frac{\partial}{\partial t}\Big[e^{\lambda t}P(n;t)\Big] = e^{\lambda t}\lambda P(n-1;t) \tag{64}$$

For $n = 1$ we have

$$\frac{\partial}{\partial t}\Big[e^{\lambda t}P(1;t)\Big] = e^{\lambda t}\lambda e^{-\lambda t} = \lambda \tag{65}$$

That is, integrating both sides and multiplying by $e^{-\lambda t}$

$$P(1;t) = \lambda t e^{-\lambda t} \tag{66}$$

For arbitrary n, we'll show by induction that

$$P(n;t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t} \tag{67}$$

We have already derived the result for $n = 0$ and for $n = 1$. We now apply induction and, assuming that the solution is valid for n, we show that it is valid for $n + 1$. We have:

$$\begin{array}{rcl} \frac{\partial}{\partial t}\Big[e^{\lambda t}P(n+1;t)\Big] & = & e^{\lambda t}\lambda P(n;t) \\ & = & e^{\lambda t}\lambda\frac{(\lambda t)^n}{n!}e^{-\lambda t} \quad \text{(induction hypothesis)} \\ & = & \lambda\frac{(\lambda t)^n}{n!} \end{array} \tag{68}$$

So, integrating

$$e^{\lambda t}P(n+1;t) = \frac{\lambda}{n!}\int(\lambda t^n)dt = \frac{(\lambda t)^{n+1}}{(n+1)!} + C \tag{69}$$

where $C = 0$ because of the initial conditions, so

$$P(n+1;t) = e^{-\lambda t}\frac{(\lambda t)^{n+1}}{(n+1)!} \tag{70}$$

(end of example)

The distribution that results from this example:

$$P_X(x) = e^{-x}\frac{x^n}{n!} \tag{71}$$

is the **Poisson distribution** that, in the example, gives us the probability that n events take place in a time x. Figure 4 shows the shape of this distribution as a function of x for various values of n.

Figure 4: The Poisson PDF for various values of n.

Example II:
The Poisson distribution can also be seen as a limiting case of the binomial distribution. If p is the probability of success in one trial, and we play N times, then $\nu = Np$ is the expected number of successful trials, as per (54). This approximation is valid for large N. In this case, we have

$$P(n;N) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n} \tag{72}$$

Taking $N \to \infty$, we have

$$\begin{aligned} P_\nu(n) &= \lim_{N\to\infty} P(n;N) \\ &= \lim_{N\to\infty} \frac{N\cdot(N-1)\cdots(N-n+1)}{n}\frac{\nu^n}{N^n}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n} \\ &= \lim_{N\to\infty} \frac{N\cdot(N-1)\cdots(N-n+1)}{N^n}\frac{\nu^n}{n!}\left(1-\frac{\nu}{N}\right)^N\left(1-\frac{\nu}{N}\right)^{-n} \\ &= 1\cdot\frac{\nu^n}{n!}e^{-\nu}\cdot 1 \\ &= \frac{\nu^n}{n!}e^{-\nu} \end{aligned} \tag{73}$$

So, once again, we find that the number of successes has a Poisson distribution.

(end of example)

The characteristic function of the distribution (71) is

$$\tilde{P}(\omega) = e^{\lambda(e^{i\omega}-1)} \tag{74}$$

from which we obtain

$$\langle X\rangle = \lambda \tag{75}$$

1.2 Functions of Random Variables

If X is a random variable on Ω, and $f : \Omega \to \Omega'$, then $Y = f(X)$ is a random variable on Ω'. Here we shallll consider, for the sake of simplicity, the case $\Omega = \Omega' = \mathbb{R}$ (all our considerations can be generalized to arbitrary contunua Ω under fairly general conditions, essentially that Ω be a metric space). In order to determine the distribution of y, we begin with a preliminary observation. For a random variable X, let $\mathbb{P}_X[x, x + \Delta x]$ the probability that the value of X falls in $[x, x + \Delta x]$. Then, for small Δx,

$$\begin{aligned} \mathbb{P}_X[x, x + \Delta x] &= P(X \leq x + \Delta x) - P(X \leq x) \\ &= \frac{\partial}{\partial x} P(X \leq x)\Delta x + O(\Delta x^2) \\ &= P_X(x)\Delta x + O(\Delta x^2) \end{aligned} \tag{76}$$

Let now f be invertible, and $g = f^{-1}$. Then

$$\begin{aligned} P_Y \Delta y &= \mathbb{P}_Y[y, y + \Delta y] \\ &= \mathbb{P}_X[g(y), g(y + \Delta y)] \\ &\approx \mathbb{P}_X\left[g(y), g(y) + \left|\frac{dg}{dy}\right| \Delta y)\right] \\ &= P_X(g(y)) \left|\frac{dg}{dy}\right| \Delta y \end{aligned} \tag{77}$$

from which we get

$$P_Y(y) = P_X(g(y)) \left|\frac{dg}{dy}\right| \tag{78}$$

Note that equivalently one could have defined

$$P_Y(y) = \int \delta(y - f(x)) P_X \, dx = \langle \delta(y - f(x)) \rangle_X \tag{79}$$

where the subscript on the average reminds us that we are taking the average with respect to the distribution of X. From this, we can determine the characteristic function of Y:

$$\begin{aligned} \tilde{P}_Y(\omega) &= \int e^{i\omega y} P_Y(y) \, dy \\ &= \int P_X(x) \Big[\int e^{i\omega y} \delta(y - f(x)) \, dy \Big] \, dx \\ &= \int e^{i\omega f(x)} P_X(x) \, dx \\ &= \langle \exp\big[i\omega f(x)\big] \rangle_X \end{aligned} \tag{80}$$

If $Y = aX$, then

$$\tilde{P}_Y(\omega) = \langle \exp\big[i\omega a X\big] \rangle_X - \tilde{P}_X(a\omega) \tag{81}$$

* * *

Consider now the sum of two random variables: $Z = X + Y$. Each value of Z can be obtained through an infinity of events: each time X takes an arbitrary value x, and y takes a value $z - x$, Z takes the same value, namely z. Summing up all these possible events we obtain

$$P_Z(z) = \int_{-\infty}^{\infty} P_X(x) P_Y(z - x)\, dx \tag{82}$$

This is known as the *convolution* of P_X and P_Y, often indicated as $P_Z = P_X * P_Y$. The properties of the Fourier transform entail that the corresponding relation between characteristic functions is

$$\tilde{P}_Z(\omega) = \tilde{P}_X(\omega)\tilde{P}_Y(\omega) \tag{83}$$

* * *

Let $Y = \{y_1, \ldots, y_n\}$ be a set of independent and identically distributed (i.i.d.) variables with cumulative distribution $\mathcal{P}_Y$ and density P_Y. Consider the function $\min(Y)$: we are interested in finding its density $P_{\min}$ and cumulative distribution $\mathcal{P}_{\min}$. We have:

$$\mathcal{P}_Y(x) = \mathbb{P}\big[\min(Y) \leq x\big] = 1 - \mathbb{P}\big[\min(Y) \geq x\big] \tag{84}$$

We have $\min(Y) \geq x$ iff we have $y_i \geq x$ for all i, that is

$$\begin{aligned} \mathcal{P}_{\min}(x) &= 1 - \mathbb{P}\Big[\forall y \in Y.y \geq x\Big] \\ &= 1 - \mathbb{P}\big[y \geq x\big]^n \\ &= 1 - \Big(1 - \mathbb{P}\big[y \leq x\big]\Big)^n \\ &= 1 - \Big(1 - \mathcal{P}_Y(x)\Big)^n \end{aligned} \tag{85}$$

The density is

$$\begin{aligned} P_{\min}(x) &= \frac{d}{dx}\mathcal{P}_{\min}(x) \\ &= n\Big(1 - \mathcal{P}_Y(x)\Big)^{n-1} \frac{d}{dx}\mathcal{P}_Y(x) \\ &= n\Big(1 - \mathcal{P}_Y(x)\Big)^{n-1} P_Y(x) \end{aligned} \tag{86}$$

For the function $\max(Y)$, working in a similar way, we have

$$\mathcal{P}_{\max}(x) = (\mathcal{P}_Y(x))^n P_{\max}(x) \quad = n(\mathcal{P}_Y(x))^{n-1} P_Y(x) \tag{87}$$

1.3 The Central Limit Theorem

The Central Limit Theorem (important enough to be granted its own acronym: CLT) is one of the fundamental results in basic probability theory and the main reason why the Gaussian distribution is so important and so common in modeling natural events. In a nutshell, the theorem tells us the following: if we take a lot of random variables, independent and identically distributed (i.i.d.), and add them up, the result will be a random variable with Gaussian distribution. So, for example, if we repeat an experiment many times and take the average of the results that we obtain (the average is, normalization apart, a sum), no matter what the characteristics of the experiment are, the resulting average will have a (more or less) Gaussian distibution.

But, ay, there's the rub! The theorem works only in the assumption that the moments of the distributions involved be finite. We shall see shortly what happens if this assumption is not satisfied.

Let $X_1, \ldots, X_n$ be a set of i.i.d. random variables with distribution P_X, zero mean, and (finite) variance σ^2. Note that $Y = \sum_i X_i$ has zero mean and variance $n\sigma^2$, while $Y = (\sum_i X_i)/n$ has zero mean and variance σ^2/n. It is therefore convenient to work with the variable

$$Z_n = \frac{1}{\sqrt{n}} \sum_i X_i \tag{88}$$

which has zero mean and variance σ^2 independently of n.

Theorem 1.2. *For any distribution P_X with finite mean and variance, and $X_1, \ldots, X_n$ i.i.d. with distribution P_X, for $n \to \infty$, we have $Z_n \to Z_\infty$, where Z_∞ is a Gaussian random variable with zero mean and variance σ^2 equal to the variance of P_X.*

Proof. Consider the first terms of the expansion of the characteristic function of P_X:

$$\tilde{P}_X(\omega) = \int e^{i\omega x} P_X(x)\, dx = 1 - \frac{1}{2}\sigma^2\omega^2 + O(\omega^3) \tag{89}$$

The characteristic function of $Y = \sum_i X_i$ is given by (83):

$$\tilde{P}_Y(\omega) = \prod_i \tilde{P}_{X_i}(\omega) = \left[\tilde{P}_X(\omega)\right]^n \tag{90}$$

(the second equality holds because the Xs have the same distribution) while (81) with $a = 1/\sqrt{n}$ gives

$$\tilde{P}_Z(\omega) = P_Y\left(\frac{\omega}{\sqrt{n}}\right) = \left[P_Y\left(\frac{\omega}{\sqrt{n}}\right)\right]^n \approx \left(1 - \frac{\sigma^2\omega^2}{2n}\right)^n \overset{n\to\infty}{\longrightarrow} \exp(-\frac{1}{2}\sigma^2\omega^2) \tag{91}$$

Finally, from (20) we have the inverse transform

$$P_Z(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{z^2}{2\sigma^2}\Big) \tag{92}$$

□

This theorem is true, in the form in which we have presented it, only for distributions X with finite mean and variance[2]. However, the key to the theorem is an invariance property of the characteristic function of the Gaussian. Consider the equality (90); we can split it up as:

$$\tilde{P}_Z(\omega; n) = \big[\tilde{P}_X(\omega)\big]^n = \big[\tilde{P}_X(\omega)\big]^{n/2}\big[\tilde{P}_X(\omega)\big]^{n/2} = \tilde{P}_Z(\omega; n/2)\tilde{P}_Z(\omega; n/2) \tag{93}$$

Taking the limit $n \to \infty$, this gives us $P_Z(\omega) = P_Z(\omega)P_Z(\omega)$. That is: the condition for a distribution to be a central limit is that the product of two characteristic functions have the same functional form as the original distributions. As we have seen in (24), the Gaussian distribution does have this property. Nay: it is the *only* distribution with finite moments that has this property, hence its appearence in the theorem in the finite moments case, and hence its great importance in application as a model of many processes resulting from the sum of identical sub-processes.

If we abandon the finite moment hypothesis, however, there is a more general distribution to which (93) applies: the stable Levy distribution. So, a more general form of the CLT can be enunciated as:

Theorem 1.3. *For any distribution P_X, and $X_1, \ldots, X_n$ i.i.d. with distribution P_X, for $n \to \infty$, we have*

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = Z_\infty \tag{94}$$

where Z_∞ is a random variable with Levy distribution. If the variance of P_X is finite and equal to σ^2, then Z_∞ has a Gaussian distribution with variance σ^2.

[2] I have assumed zero mean since, if the mean of the X is non-zero, the mean of Z goes to infinity; this doesn't represnt a major hurdle for the theorem, which can easily be generalized by subtracting the mean from the variables X and then adding it back.

2 Stochastic Processes

A **Stochastic process** is a collection of random variables $X(t)$ (often written as X_t) indexed by a parameter t, usually identified with time. Normally, we assume $t \in \mathbb{N}$ (in which case we talk about a *discrete time* stochastic process) or $t \in \mathbb{R}^+$ (*continuous time* stochastic process). Here we shall deal mainly with the first type, although many of the definitions in this section apply to both. The instantiation of a random variable is a value $x \in M$ (viz., the value of the variable that we get if one of the outcomes $\omega \in \Omega$ obtains); the instantiation of a (dicrete time) stochastic process is a **trajectory** $x : \mathbb{N} \to M$.

Stochastic process

trajectory

Let $P(X(t) = x)$ (or simply $P(x;t)$) the probability that X take value x at time t. Let $x \in M$: if M is discrete, then $P(x;t)$ is a probability, while if M is continuous, it is a probability density. Similarly, $P(X(t_1) = x_1, X(t_2) = x_2)$ (or $P(x_1, x_2; t_1, t_2)$) is the probability that $X(t_1)$ take value x_1 and $X(t_2)$ take value x_2. The generalization to k values

$$P(x_1, \ldots, x_k; t_1, \ldots, t_k) \tag{95}$$

gives complete information about the process. Needless to say, these (in principle, infinite) values are quite uncomfortable to calculate. We need to find easier ways to get at least some information about the process.

Example III:
As a first example, let us consider the throw of a die. We'd better not be bothered with averages, so we shall subtract a suitable value to the result of the throw so as to obtain a distribution with zero mean. To simplify things, we shall stay with integer numbers, which means that we can't use a standard, six-face die (which has mean score of 3.5). So, we shall use a five-face die instead and subtract the average score (3) from the result. This way, each throw will have a result in $\{-2, \ldots, 2\}$ with a nice zero mean[3]. In this case

$$P_x(m,t) = \begin{cases} \frac{1}{5} & m \in \{-2, \ldots, 2\} \\ 0 & \text{otherwise} \end{cases} \tag{96}$$

and, since the throws are independent and have no influence on each other,

$$P(m_1, m_2; t_1, t_2) = P(m_1; t_1)P(m_2; t_2) = \begin{cases} \frac{1}{25} & (m_1, m_2) \in \{-2, \ldots, 2\} \times \{-2, \ldots, 2\} \\ 0 & \text{otherwise} \end{cases} \tag{97}$$

[3]It is impossible to physically build a five-face die: there is no regular polyhedron with five faces. One way to do it is to use a regular six-face die and simply ignore the six. That is, if we throw a six, we repeat the throw as many time as it is necessary to obtain something other than a six. **Excercise:** show that this is equivalent to throwing a five-face fair die. Intuitively the result is obvious, but it is nice to prove it. There are at least two ways: one doing the actual calculations, the other using symmmetry.

and

$$P(m_1,\ldots,m_k;t_1,\ldots,t_k)=\begin{cases}\left(\frac{1}{5}\right)^k & (m_1,\ldots,m_k)\in\{-2,\ldots,2\}^k\\ 0 & \text{otherwise}\end{cases} \tag{98}$$

Here are the beginnings of two trajectories of the process (yes: I actually did throw the die)

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$X_t^{(1)}$	1	1	0	0	-2	2	0	2	2	0	1	-1	2	2	1	1	2	-2	0	1
$X_t^{(2)}$	1	1	0	-1	-1	-1	-2	1	0	-2	-1	-2	2	1	-2	2	2	-2	-2	2

(end of example)

Example IV:

Consider the case in which we still throw the die and correct for average, as in the previous example. This time, however, we want to know the *sum* of all the scored obtained up to time t. That is, if X_t is the process of the previous example, we are interested in the process

$$Y_t=\sum_{\tau=0}^{t}X_\tau \tag{99}$$

For the two trajectories of X_t of the previous example, we obtain the corresponding trajectories of Y as

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$Y_t^{(1)}$	1	2	2	2	0	2	2	4	6	6	7	6	8	10	11	12	14	12	12	13
$Y_t^{(2)}$	1	2	2	1	0	-1	-3	-2	-2	-4	-5	-7	-5	-4	-6	-4	-2	-4	-6	-4

The distribution $P_Y(m;t)$ is no longer constant in time. The probability depends on the distribution of X (which is constant) as well as on $P_Y(m;t-1)$ (which is not) as

$$P_Y(m;t)=\sum_k P_X(k)P_Y(m-1;t-1)=\frac{1}{5}\sum_{k=-2}^{2}P_Y(m-k;t-1) \tag{100}$$

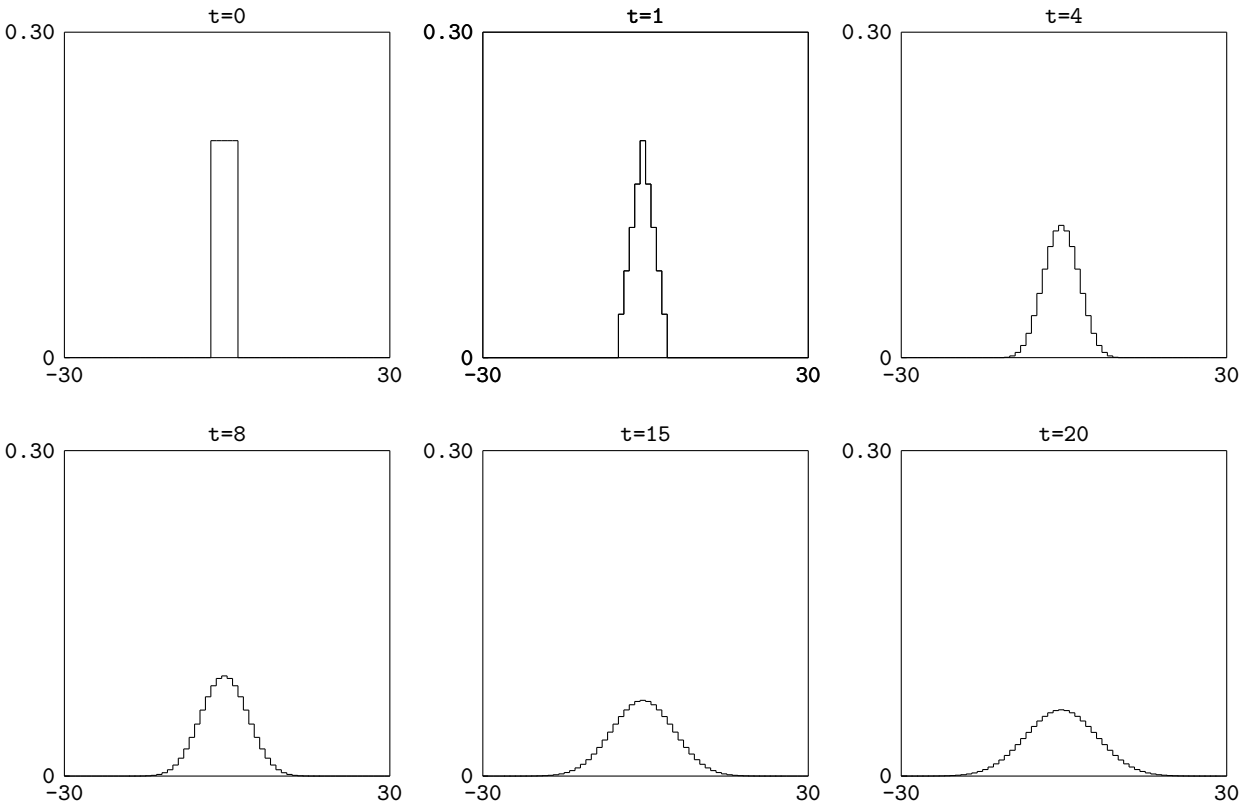

Figure 5: The distribution of Y (the cumulative score of a series of zero-average die throws) is not constant in time. These are the distributions for various values of t.

The distribution, for various values of t is shown in Figure 5 For $t \to \infty$, as per the central limit theorem, the distribution will tend to a Gaussian. However, lacking the normalization term $1/\sqrt{t}$ that appears in the theorem, the variance will diverge: $\sigma^2 \to \infty$. The distributions $P_Y(m_1, m_2; t_1, t_2)$ are in general quite complex to compute, as the value of m_1 at time t_1 depends on the value m_2 at time t_2 through all intervening values. For $t_2 = t_1 - 1$ we have

$$P_Y(m, m'; t, t-1) = P_Y(m'; t-1) P_Y(m; t|m'; t-1) \tag{101}$$

The second factor is the uniform distribution centered around m', so

$$P_Y(m, m'; t, t-1) = \frac{1}{5} \sum_{m=m'-1}^{m'+1} P_Y(m; t) \tag{102}$$

(end of example)

Example V:

Consier again the process X_t but, this time, define the process Z_t by adding up the last two samples of X_t:

$$Z_t = X_t + X_{t-1} \tag{103}$$

(assume $X_{-1} = 0$, so as to be able to define Z_0). In this case, the trajectories that correspond to the given trajectories of X_t are

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$Z_t^{(1)}$	1	2	1	0	-2	0	2	2	4	2	1	0	1	4	3	2	3	0	-2	1
$Z_t^{(2)}$	1	2	1	-1	-2	-2	-3	-1	1	-2	-3	-3	0	3	-1	0	4	0	-4	0

The distribution $P_Z(m)$ is that of the sum of two uniform distributions, and can be seen to be

$$P_Z(m) = \begin{cases} \frac{1}{25}(5 - |m|) & |m| \leq 4 \\ 0 & \text{otherwise} \end{cases} \tag{104}$$

The distribution $P_Z(m, m'; t, t-1)$ is a bit more complicated. We have

$$P_Z(m, m'; t, t-1) = P_Z(m'; t-1) P_Z(m; t|m'; t-1) \tag{105}$$

The first term is easy: it is (104), the second is a tad more problematic. If we knew the value X_{t-1}, things would be easy: $P_Z(Z_t = m|X_{t-1} = k)$ would just be the uniform distribution centered around k, that is, $P_X(m-k)$. But we don't know X_{t-1}, we only know Z_{t-1}. We can, however, sum over all probabilities of X_{t-1} conditioned to Z_{t-1} being equal to m':

$$\begin{aligned} P(Z_t = m|Z_{t-1} = m') &= \sum_k P(Z_t = m|X_{t-1} = k) P(X_{t-1} = k|Z_{t-1} = m;) \\ &= \sum_k P_X(m-k) P(X_{t-1} = k|Z_{t-1} = m;) \end{aligned} \tag{106}$$

on the other hand, we have

$$P(X_{t-1} = k|Z_{t-1} = m;) = P(X_{t-1} = k|X_{t-1} + X_{t-2} = m;) = P(X_{t-2} = m' - k) = P_X(m' - k) \tag{107}$$

So

$$P(m; t|m'; t-1) = \sum_k P_X(m-k) P_X(m'-k) \tag{108}$$

and

$$P_Z(m, m'; t, t-1) = P_Z(m') \sum_k P_X(m-k) P_X(m'-k) \tag{109}$$

(end of example)

For fixed t, X_t is a stochastic variable with distribution $P_t(x) = P(x; t)$. We can characterize this distribution using pretty much the same quantities that we used for stochastic variables, with the difference that, in this case, these quantities will be functions of time. We shall, for the sake of simplicity, consider stochastic processes in which the range is $\mathbb{R}$. The **mean function** $\mu : \mathbb{N} \to \mathbb{R}$ is defined as

mean function

$$\mu_t = \mathbb{E}[X_t] \tag{110}$$

The **variance** of the process is the function $\gamma : \mathbb{N}^2 \to \mathbb{R}$ defined as

variance

$$\gamma(t, t) = \mathbb{E}\big[(X_t - \mu_t)^2\big] \tag{111}$$

You are probably wondering why the parameter t is repeated. This is so because the variance is a special case of the **covariance function**

covariance

$$\gamma(s, t) = \mathbb{E}\big[(X_s - \mu_s)(X_t - \mu_t)\big] \tag{112}$$

From this function we derive the **autocorrelation function** $\rho : \mathbb{N}^2 \to \mathbb{R}$ defined as

autocorrelation

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{113}$$

from the inequality $|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t)$ we have

$$-1 \leq \rho(s,t) \leq 1 \tag{114}$$

* * *

One problem with the study of stochastic processes is that in general the distribution $P_t(x)$ is a function of t so, in order to understand the process, we must characterize a different distribution for each t. In general this is impossible, unless we can explicit write down the functional form of this dependency. The simplest (and one of the most useful) of these functional forms is the constant; we say that a process is **stationary** if $P_t(x) = P(x)$, that is, if the distribution of the process is independent of time. A process that satisfies this condition is referred to as **strongly** stationary.

stationary process

Definition 2.1. *A stochastic process X_t with distribution $P(x;t)$ is **strongly stationary** if, for all x, t_1, t_2, it is $P(x;t_1) = P(x;t_2)$.*

strongly stationary

An equivalent condition is that, for all $x_1, x_2, t_1, t)2, \tau$,

$$P(x_1, x_2; t_1, t_2) = P(x_1, x_2; t_1 + \tau, t_2 + \tau) \tag{115}$$

This is a fairly strong condition, stronger than we need in most cases. For most purposes, we can weaken it.

Definition 2.2. *A stochastic process X_t with distribution $P(x;t)$ is **weakly stationary** if, for all t, s, τ it is*

weakly stationary

$$\mathbb{E}\big[X_t\big] = \mu \tag{116}$$

$$\mathbb{E}\big[X_t^2\big] < \infty \tag{117}$$

$$\gamma(s,t) = \gamma(s+\tau, t+\tau) \tag{118}$$

Note that (115) implies (118), which is therefore true for strongly stationary processes, while (118) and (116) imply that, for all s, t,

$$\mathbb{E}\big[X_t^2\big] = \mathbb{E}\big[X_s^2\big] \tag{119}$$

and, together with (118),

$$\mathbb{E}\big[(X_t - \mu)^2\big] = \sigma^2 \tag{120}$$

Condition (118) can be written as

$$\gamma(s,t) = \gamma(s-t) \tag{121}$$

That is, the covariance of X_t and X_s depends only on their distance (with sign) in time, and not on the absolute time at which they are considered.

Definition 2.3. *A stochastic process X_t is* ***white*** *if $\gamma(s,t) = \delta_{s,t}\sigma^2$*

white process

Normally, for us, a white (stationary) process will be a process w_t with $\mu_t = 0$ (the average, being a constant, is not very interesting: we can always substract it, study the process with zero average, and then add it back), and $\mathbb{E}[w_t^2] = \sigma^2$, we shall write it as $w \sim wN(0, \sigma^2)$.

That a process is white means, essentially, that the past gives un no information about the present value of w_t. This doesn't mean that we can't make a prediction on w_t: the distribution $wN(0, \sigma^2)$ may very well allow us to make good predictions. What is means is that it doesn't matter whether we know the values $w_0, \ldots, w_{t-1}$ or not, our predictions and their accuracy will not change. We can write:

$$P(x_t; t | x_0, \ldots, x_{t-1}; 0, \ldots, t-1) = P(x_t; t) \tag{122}$$

Example VI:

The process X_t that we have defined above has the same distribution $P_X(m)$ independently of time. It is, therefore, strongly stationary, and

$$\begin{aligned} &\mathbb{E}[X_t] = 0 \\ &\mathbb{E}[X_t^2] = \frac{1}{5}\sum_{m=-2}^{2} m^2 = 2 \\ &\gamma(s,t) = 2\delta_{s,t} \end{aligned} \tag{123}$$

The process is white (the results of one throw are independent of the results of the others).

(end of example)

Example VII:

The process Y_t has a distribution that depends on t, it is therefore not strongly stationary. It has $\mathbb{E}[X_t] = 0$ (constant), but $\mathbb{E}[X_t^2] \sim \sigma^2 t$, therefore, by virtue of (120) it is not weakly stationary. For the correlation we have (we can assume $t > s$, the other case being equal for symmetry)

$$\begin{aligned}\gamma(t,s) &= \mathbb{E}[Y_t Y_s] \\ &= \sum_{m_1}\sum_{m_2} m_1 m_2 P(Y_t = m_1, Y_s = m_2) \\ &= \sum_{m_1}\sum_{m_2} m_1 m_2 P(Y_s = m_2) P(Y_t = m_1 | Y_s = m_2)\end{aligned} \tag{124}$$

If $s = t-1$,

$$\begin{aligned}\gamma(t,t-1) &= \mathbb{E}[Y_t Y_{t-1}] \\ &= \sum_{m_1}\sum_{m_2} m_1 m_2 P(Y_t = m_1, Y_{t-1} = m_2) \\ &= \sum_{m_1}\sum_{m_2} m_1 m_2 P(Y_{t-1} = m_2) P(Y_t = m_1 | Y_{t-1} = m_2) \\ &= \frac{1}{5}\sum_{m_2}\sum_{m_1=m_2-2}^{m_2+2} m_1 m_2 P_Y(m_2, t-1)\end{aligned} \tag{125}$$

The process is not white, and condition (121) does not hold.

(end of example)

Example VIII:

The process Z_t has a distribution (104) independent of t, it is therefore strongly stationary. We

have

$$\begin{aligned}\mathbb{E}[Z_t] &= 0 \\ \mathbb{E}[Z_t^2] &= \sum_m m^2 P_Z(m) = \frac{1}{25}\sum_{m=-2}^{2} m^2(5-|m|) \\ &= \frac{2}{25}\sum_{m=0}^{2} m^2(5-m) \\ &= 4\end{aligned} \tag{126}$$

The value Z_t depends only on X_t and X_{t-1}, and these are independent of the previous values of X, therefore

$$\gamma(t, t-\tau) = 0 \quad \tau \geq 2 \tag{127}$$

For $\tau = 1$ we have

$$\begin{aligned}\gamma(t, t-1) &= \sum_{m_1}\sum_{m_2} m_1 m_2 P(Z_{t-1} = m_2) P(Z_t = m_1 | Z_{t-1} = m_2) \\ &= \sum_{m_1}\sum_{m_2} m_1 m_2 P_Z(m_2) \sum_k P_X(m_1 - k) P_X(m_2 - k)\end{aligned} \tag{128}$$

where we have applied (109). Doing the calculations (do them! it is a good exercise in manipulating interdependent indices and the associated conditions) one finds

$$\gamma(t, t-1) = \frac{7}{5} = 1.4 \tag{129}$$

(end of example)

In a stochastic stationary process, we have two ways of computing averages: we can compute the *ensemble average* $\langle X(t) \rangle$, that is, the average of the random variable $X(t)$, or the mean value along a trajectory

$$\bar{X} = \lim_{T\to\infty} \frac{1}{T}\int_0^T x(t)\, dt \tag{130}$$

A process is *ergodic* if the two coincide

$$\langle X \rangle = \bar{X} \tag{131}$$

Ergodicity is an important property for us: many times we shall be interested in the characteristics of the motion of one individual, but many of the equations that we shall use involve ensemble probabilities based on a whole population. Ergodicity allows us to switch from one to the other with impunity.

2.1 Gaussian processes

A stochastic process $X(t)$ is *Gaussian* with zero mean if $\mathbb{E}[X(t)] = \mu_t = 0$ and

$$P(x_i, t_i) = \sqrt{\frac{A_{ii}}{2\pi}} \exp\Big(-\frac{1}{2} A_{ii} x^2\Big) \tag{132}$$

($A_{ii} > 0$). The joint probability $P(x_1, t_1; \ldots; x_n, t_n)$ then follows a multivariate Gaussian distribution

$$P(x_1, t_1; \ldots; x_n, t_n) = \frac{\det(\mathbf{A})^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i,j=1}^{n} x_i A_{ij} x_j\right] \tag{133}$$

Where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric (strictly) positive definite. The matrix $\mathbf{A}$ is a measure of the covariance between two variables of the Gaussian process

$$\mathbb{E}[X(s)X(t)] = \gamma(s,t) = (\mathbf{A}^{-1})_{st} \tag{134}$$

(this is true since we assume zero mean). A process is uncorrelated if $\mathbf{A}$ is diagonal, then

$$\gamma(s,t) = A_{st}^{-1} \delta_{s,t} \tag{135}$$

and the process is white.

2.2 Wiener Processes

A *Wiener process* W is a process in which the variables $W(t)$ are real and with independent increments $W(t_2) - W(t_1)$ that follow a Gaussian distribution. That is, they define a conditional probability

$$P(w_2, t_2 | w_1, t_1) = \frac{1}{\sigma\sqrt{2\pi(t_2 - t_1)}} \exp\left[-\frac{(w_2 - w_1)^2}{2\sigma^2(t_2 - t_1)}\right] \tag{136}$$

from which the covariance can be computed

$$\begin{aligned} \langle (W(t_2) - \langle W\rangle)(W(t_1) - \langle W\rangle)\rangle &= \langle (W(t_2) - W(0))(W(t_1) - W(0))\rangle \\ &= \int_{-\infty}^{\infty} (w_2 - w_0)\, dw_2 \int_{-\infty}^{\infty} dw_1 (w_1 - w_0) P(w_2, t_2; w_1, t_1) \\ &= \sigma^2 \min(t_1, t_2) + w_0^2 \end{aligned} \tag{137}$$

Note that this entails that a Wiener process is neither white nor stationary. From this we get

$$\langle W(t)^2 \rangle = \sigma^2 t + w_0^2 \tag{138}$$

Wiener processes are related to Gaussian processes, in particular to uncorrelated (white) Gaussian processes. Let $X(t)$ be a Gaussian process with $\langle X(t_1)X(t_2)\rangle = \sigma^2\delta(t_2 - t_1)$, and define a new stochastic process as the integral of $X(t)$:

$$Y(t) = \int_0^t X(u)\,du \tag{139}$$

then

$$\begin{aligned}\langle Y(t_2)Y(t_1)\rangle &= \int_0^{t_2} du_2 \int_0^{t_1} du_1 \langle X(u_1)X(u_2)\rangle \\ &= \int_0^{t_2} du_2 \int_0^{t_1} du_1 \delta(u_2 - u_1)\end{aligned} \tag{140}$$

By the properties of the Dirac function

$$\int_0^{t_1} du_1 \delta(u_2 - u_1) = \begin{cases} 1 & 0 < u_2 < t_1 \\ 0 & \text{otherwise} \end{cases} \tag{141}$$

Then

$$\langle Y(t_2)Y_(t_1)\rangle = \sigma^2 \min(t_2, t_1) \tag{142}$$

which coincides with (138) for $w_1 = 0$. That is, the integral of a Gaussian process is a Wiener process.