Contents

A quick introduction to probability theory

Probability theory focuses on the results of one or many experiments which outcomes cannot be predicted deterministically. A certain number of outcomes constitute what is called an event. It is not necessary to specify the physical nature of the outcomes. For example, they could be the "head" or the "tail" occurrence in a coin toss experiment. In some cases, these outcomes can be characterized by real numbers like "the time instant of a phone call" or "the position coordinate of a particle of pollen in a fluid suspension". However, the numerical character of a result is just one of the possible ways of labeling or identifying it.

0.1 Basic notions

In order to obtain a precise definition of the "probability concept", probability theory introduces a mathematical structure, named *probability space*, which is in fact a collection of three different objects:

1. An abstract space Ω, called *sample space* which contains all distinguishable *elementary outcomes* or results of an experiment. These elementary outcomes might be names, numbers, complicated signals,. . . .

2. An *event space* (*sigma-field* or *sigma-algebra*) $\mathcal{F}$ consisting of the collection of subsets of the sample space Ω we wish to consider as possible events and to which we would like to a probability. Every element in $\mathcal{F}$ is called an *event.* The event space is required to have a closed algebraic structure in the following sense: Any finite or countably infinite sequence of the basic set operations (union, intersection, complementation, difference, symmetric difference) on events must produce other events contained in $\mathcal{F}$.

3. A *probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$, i.e. an assignment of a number between 0 and 1 to every event. A probability measure must obey certain rules or *axioms* and will be practically computed by integrating or summing its associated *probability density.*

The necessity of the introduction the event space in place of dealing directly with the individual points of the sample space can be understood on the basis of the following practical example. If we spin an unbiased pointer around its rotation pin, the outcome is known to be equally likely any number between 0 and 2π. Then, the probability that any particular point such as 0.371529613 or exactly $1/\pi$ occurs is zero, because within the interval $[0, 2\pi)$ there is an uncountable infinity of possible numbers, none more likely than the others (the set of both irrational and rational numbers in an interval is uncountable). Hence, knowing only that the probability of each and every point is zero, is not useful for making inferences about the probabilities of other events such as, for example, the outcome being between $1/3$ and $3/4$ or between $\pi/2$ and π.

0.1.1 Probability spaces

Here we introduce more formal definitions of the mathematical structure of probability spaces. The basic idea is to relate any probability space to a model space in which explicit calculations are easily performed.

Definition (Sample space Ω)**.** A *sample space* Ω is an abstract space, a nonempty collection of points or members or elements called *sample points* (or *elementary events* or *elementary outcomes*).

Definition (Event space or σ-algebra $\mathcal{F}$)**.** A σ-algebra (or *sigma field*) $\mathcal{F}$ is a collection of subsets of Ω such that:

(i) $\emptyset \in \mathcal{F}$;

(ii) If $A \in \mathcal{F}$ then $A^c \equiv \Omega - A \in \mathcal{F}$ (*closed under complement*);

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$ (*closed under countable unions*).

In short, a σ-algebra is a collection of subsets which must be "closed" with respect to the basic set operations.

There are two extreme examples of σ-algebras:

- The collection $\{\emptyset, \Omega\}$ is a σ-algebra of subsets of Ω;
- The set $S(\Omega)$ of *all* subsets of Ω is a σ-algebra

Clearly any σ-algebra $\mathcal{F}$ of subsets of Ω lies between these two extremes:

$$\{\emptyset, \Omega\} \subset \mathcal{F} \subset S(\Omega) \tag{1}$$

Some properties of a σ-algebra are the following:

Proposition 0.1.1. Let $\mathcal{F}$ be a σ-algebra of subsets of Ω then

1. $\Omega \in \mathcal{F}$;
2. If $A_1, \cdots, A_n \in \mathcal{F}$, then $A_1 \cup \cdots \cup A_n \in \mathcal{F}$;
3. If $A_1, \cdots, A_n \in \mathcal{F}$, then $A_1 \cap \cdots \cap A_n \in \mathcal{F}$
4. If $A_1, A_2, \cdots$ is a countable collection of sets in $\mathcal{F}$, then $\cap_{n=1}^{\infty} A_n \in \mathcal{F}$;
5. If $A, B \in \mathcal{F}$, then $A - B \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ constitute a *measurable space*, a space on which a measure can be defined. If the measure satisfies some specific properties, then it defines a probability measure on this space.

Definition (Probability measure $\mathbb{P}$)**.** Let $\mathcal{F}$ be a σ-algebra of subsets of Ω. A probability measure on a measurable space $(\Omega, \mathcal{F})$ is an application $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

(i) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$;

(ii) If $A_1, A_2, \ldots \in \mathcal{F}$ then $\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k)$ (subadditivity);

(iii) If $A_1, A_2, \ldots$ are *disjoint* sets in $\mathcal{F}$ then $\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.

Putting all together we have the following definition of the probability space:

Definition (Probability space $(\Omega, \mathcal{F}, \mathbb{P})$)**.** A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space provided Ω is any set, $\mathcal{F}$ a σ-algebra of subsets of Ω, and $\mathbb{P}$ is a probability measure on $\mathcal{F}$. A set $A \in \mathcal{F}$ is called an *event*; the points $\omega \in \Omega$ are called *sample points*.

There is a simple σ-algebra which allows for the construction of a probability space which serves as a model for all other spaces.

Definition (Borel σ-algebra $\mathcal{B}$)**.** $\mathcal{B}$ is the smallest σ-algebra containing all the open subsets of $\mathbb{R}$.

Given any nonnegative integrable function p such that $\int_{\mathbb{R}} dxp(x) = 1$, consider the probability measure on $\mathcal{B}$ given by

$$\mathbb{P}(B) \equiv \int_B dxp(x) \tag{2}$$

for each $B \in \mathcal{B}$. The triple $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ is a probability space. We call p the *density* of the probability measure $\mathbb{P}$. In the model-space $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ the events are identified through real numbers. Loosely speaking, the probability density function p establishes which numbers are more or less probable as an outcome. The importance of the model relies on the fact that since the probability is defined through an ordinary integral, it allows explicit calculations. Any abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is connected to the above model through mappings called *random variables.*

Definition (Random variable X)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A mapping

$$X : \Omega \to \mathbb{R}$$

is called a random variable if for each $B \in \mathcal{B}$

$$X^{-1}(B) \in \mathcal{F}.$$

Remark. Following common notations in probability theory, we usually write X in place of $X(\omega)$. In the same spirit, $\mathbb{P}(X^{-1}(B))$ is indicated as $\mathbb{P}(X \in B)$. More specifically,

$$\mathbb{P}(X \in B) \equiv \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right) \tag{3}$$

Analogously, given $x \in \mathbb{R}$ the notation $\mathbb{P}(X \leq x)$ means the probability measure of the event $A \in \mathcal{F}$ defined by the property $\forall y \in X(A)$, $y \leq x$:

$$\mathbb{P}(X \leq x) \equiv \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq x\}\right) \tag{4}$$

Definition (Probability distribution function P_X)**.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, its probability *distribution* function is the function $P_X : \mathbb{R} \to [0, 1]$ defined by:

$$P_X(x) \equiv \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}. \tag{5}$$

It is often useful to work with the density distribution functions.

Definition (Probability density function p_X)**.** Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, its probability *density* function is the function $p_X : \mathbb{R} \to \mathbb{R}^+$ such as:

$$P_X(x) = \int_{-\infty}^{x} dx' p_X(x') \quad \forall x \in \mathbb{R}. \tag{6}$$

Of course, we have

$$\mathbb{P}(x_1 \leq X \leq x_2) = \textstyle\int_{x_1}^{x_2} dx' p_X(x'), \tag{7}$$

$$p_X(x)dx = \mathbb{P}(x \leq X \leq x + dx). \tag{8}$$

Remark (Normalization)**.** It is straightforward to see that

$$P_X(+\infty) - P_X(-\infty) = \int_{\mathbb{R}} dx' p_X(x') = 1. \tag{9}$$

Note. For *discrete* random variables all the above holds if one considers

$$p_X(x) = \sum_i p_i \delta(x - x_i), \tag{10}$$

where $p_i \equiv \mathbb{P}(X = x_i)$.

0.2 Moments of a random variable

Definition (Expected value $\mathbb{E}\{X\} = \eta$)**.** The expected value (or mean value) of the random variable X is the real number:

$$\boxed{\mathbb{E}\{X\} = \int_{-\infty}^{\infty} dx\; x\; p_X(x) \equiv \eta.} \tag{11}$$

Definition (Variance σ^2)**.** The variance of the random variable X is the real number:

$$\boxed{\sigma^2 \equiv \mathbb{E}\{(X-\eta)^2\} = \int_{-\infty}^{\infty} dx (x-\eta)^2 p_X(x).} \tag{12}$$

Definition (Moment of order r, m_r)**.** The moment of order $r = 1, 2, \ldots$ of the random variable X is the real number:

$$m_r \equiv \mathbb{E}\{(X)^r\} = \int_{-\infty}^{\infty} dx\; x^r p_X(x). \tag{13}$$

Definition (Central moment of order r, μ_r)**.** The central moment of order $r = 1, 2, \ldots$ of the random variable X is the real number:

$$\mu_r \equiv \mathbb{E}\{(X-\eta)^r\} = \int_{-\infty}^{\infty} dx (x-\eta)^r p_X(x). \tag{14}$$

0.3 Examples of probability distributions

Normal distribution $\mathbb{N}(\eta, \sigma^2)$
One of the most important real valued random variables is the Gaussian or normal distribution. A random variable is Gaussian with mean value η and variance σ^2 if its probability density function is given by

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\eta)^2}{2\sigma^2}}. \tag{15}$$

We also indicate $X \in \mathbb{N}(\eta, \sigma^2)$

Uniform distribution
One says that a random variable X is uniformly distributed within the interval (x_1, x_2) if its probability density is defined as

$$p_X(x) = \begin{cases} \frac{1}{x_2 - x_1} & x_1 \le x \le x_2 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

For example the built in routines for the generation of random numbers in computers generate random numbers uniformly distributed between 0 and 1.

Cauchy distribution
One says that a random variable X follows a Cauchy distribution if its probability density is defined as

$$p_X(x) = \frac{\alpha/\pi}{\alpha^2 + (x - x_0)^2} \tag{17}$$

where x_0 is the *location parameter*, giving the location of the peak of the distribution and α is the *scale parameter* which specifies the half-width at half-maximum. As a probability distribution, it is known as the *Cauchy distribution* while among physicists it is also known as the *Lorentz distribution* or the *Breit-Wigner* distribution. Its importance in physics is largely due to the fact that it is the solution of the differential equation characterizing *forced resonance*. In *spectroscopy*,

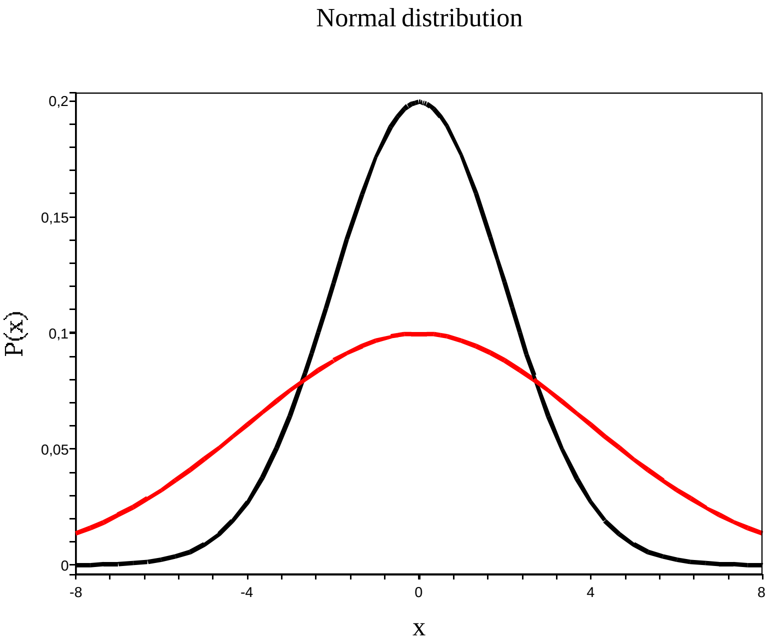

Figure 1: Two example of Normal distributions. Both distribution have mean zero but different σ^2. In particular the lower curve has $\sigma^2 = 4$ while the upper curve has $\sigma^2 = 2$.

Figure 2: Two example of uniform distributions.

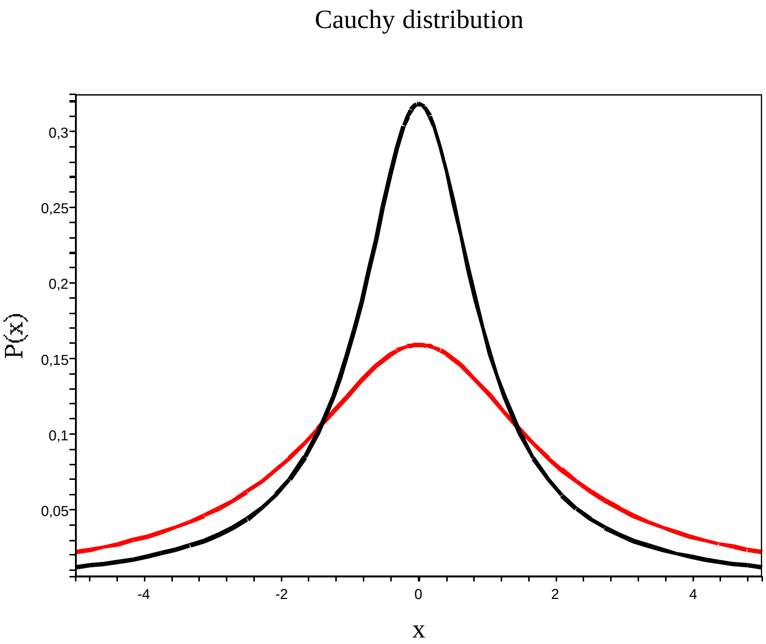

Figure 3: Cauchy distributions (upper $\alpha = 1$, lower $\alpha = 2$). In these two examples $x_0 = 0$. The maximum is indeed at $x = 0$ and has the value $\frac{1}{\alpha\pi}$

it describes the line shape of spectral lines which are broadened by several mechanisms including resonance broadening. The special case when $x_0 = 0$ and $\alpha = 1$ is called the *standard Cauchy distribution*. Note that the mean of the Cauchy distribution is undefined while the second moment and higher moments are divergent.

Binomial distribution

An example of a discrete random variable is given by the binomial distribution for which the probability that the random variable assumes the value k among n possible values is given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n, \quad p + q = 1. \tag{18}$$

Its probability density function is then

$$p_X(x) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \delta(x - k). \tag{19}$$

A practical example of a random variable following the binomial distribution comes from the coin toss experiment. The outcomes are a series $\zeta_1, \zeta_2, \dots, \zeta_n$ of "heads" and "tails". Let us define X such that $X(\zeta_1, \zeta_2, \dots, \zeta_n)$ is equal to the number of "heads":

$$X(\zeta_1, \zeta_2, \cdots, \zeta_n) = k \qquad \text{if } k \leq n \text{ are the "head" outcomes.} \tag{20}$$

In this way $\{X = k\}$ is the event "k heads in n outcomes". Since each toss is an independent process the probability of a given string having k heads is given by $p^k q^{n-k}$ where p is the probability that in a given toss the coin gives head and q is the probability that in a given toss the coin gives tail.

On the other hand the total number of strings having k heads out of n slots is given by the combination $C_k^n = \binom{n}{k}$. The probability of having k heads out of n trails is then given by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad \text{with} \quad p = q = 1/2. \tag{21}$$

Poisson distribution

A random variable X follows a Poisson distribution with parameter $\lambda(>0)$ is the probability that the value k occurs is given by

$$\mathbb{P}(X=k;\lambda)=e^{-\lambda}\frac{\lambda^k}{k!},\qquad k=0,1,\ldots \tag{22}$$

The probability density is given by

$$p_{X;\lambda}(x)=\sum_{k=0}^{n}e^{-\lambda}\frac{\lambda^k}{k!}\delta(x-k) \tag{23}$$

It is easy to show that (Exercise)

$$\mathbb{E}\{X;\lambda\}=\lambda,\qquad \sigma_X^2=\lambda. \tag{24}$$

The Poisson distribution has many applications since it can be used as a good approximation of a binomial r.v. with parameters (n,p) when n is big enough and p is sufficiently small in such a way that the term np as a finite value (see Problem) Some examples of random variables following the Poisson distribution are

- The number of phone calls reaching a node at different time intervals.
- The number of missprints in a page of a book.
- The number of people that enter, say into a post office, at different time intervals.

0.4 Conditional probabilities

In the axiomatic theory of probability if $B\in\mathcal{F}$ is an event such that $\mathbb{P}(B)\neq 0$, then the probability of a given event $A\in\mathcal{F}$ *conditioned* by B is defined as

$$\mathbb{P}(A|B)\equiv\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}. \tag{25}$$

where $\mathbb{P}(A\cap B)$ is the probability of the both events A and B occur.

Definition (Conditional distribution function)**.** The probability distribution $P_X(x|B)$ of the random variable X, conditioned by B is defined as the conditional probability of the event $A\equiv\{X\leq x\}$:

$$P_X(x|B)\equiv\mathbb{P}\{X\leq x|B\}=\frac{\mathbb{P}\{X\leq x\cap B\}}{\mathbb{P}(B)}. \tag{26}$$

If the event B is expressed in terms of X then $P_X(x|B)$ can be determined in terms of $P_X(x)$.

0.5 Characteristic functions

Characteristic functions are convenient tools in probability theory. They are the Fourier transforms of the probability density functions and often offer a practical alternative for calculating interesting quantities.

Definition (Characteristic function, f_X)**.** The characteristic function of a random variable X is the complex function $f_X:\mathbb{R}\to\mathbb{C}$:

$$f_X(k)\equiv\mathbb{E}\{e^{ikx}\}=\int_{-\infty}^{+\infty}dx\;e^{ikx}\;p_X(x). \tag{27}$$

The characteristic function $f_X(k)$ uniquely identifies the probability density function of the random variable. In fact, from the characteristic function one can re-obtain $p_X(x)$ through an inverse Fourier transform:

$$p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dx\, e^{-ikx}\, f_X(k). \tag{28}$$

From the normalization condition on $p_X(x)$, the characteristic function inherits the following property

$$f_X(0) = 1. \tag{29}$$

Moreover,

$$|f_X(k)| \leq 1 \qquad \forall k \in \mathbb{R}, \tag{30}$$

$$f_X(-k) = f_X(k)^* \qquad \forall k \in \mathbb{R}. \tag{31}$$

Finally, if p_X is even ($p_X(-x) = p_X(x)$), then f_X is real and even ($f_X(-k) = f_X(k)^* = f_X(k)$).

The moments m_r are easily calculated from the characteristic function:

$$\boxed{m_r = \mathbb{E}\{X^r\} = \lim_{k\to 0}(-i)^r \frac{d^r f_X(k)}{dk^r}.} \tag{32}$$

Therefore, in general a regular characteristic function can be written as the series:

$$\begin{aligned} f_X(k) &= \sum_{r=0}^{\infty} \frac{\frac{d^r f_X(k)}{dk^r}|_{k=0}}{r!}\, k^r \\ &= \sum_{r=0}^{\infty} \frac{i^r\, m_r}{r!}\, k^r \qquad (33) \\ &= 1 + i\eta k - \frac{m_2 k^2}{2} + \dots . \qquad (34) \end{aligned}$$

Examples: (see Problems)

- The characteristic function of a normal distribution $\mathbb{N}(\eta, \sigma^2)$ is

$$f_X(k) = e^{ik\eta} e^{-\frac{\sigma^2 k^2}{2}}, \quad k \in \mathbb{R}. \tag{35}$$

- The characteristic function of a Cauchy distribution is

$$f_X(k) = e^{ikx_0} e^{-\alpha|k|}, \quad k \in \mathbb{R}. \tag{36}$$

- The characteristic function of a uniform distribution is

$$f_X(k) = \frac{e^{ikx_2} - e^{ikx_1}}{ik(x_2 - x_1)}, \quad k \in \mathbb{R}. \tag{37}$$

- The characteristic function of a binomial distribution is

$$f_X(k) = \left(q + pe^{ik}\right)^n, \quad k \in \mathbb{R}. \tag{38}$$

0.6 Function of a random variable

Given a random variable X_1, consider the new random variable

$$X_2 \equiv h(X_1), \tag{39}$$

where $h:\mathbb{R}\to\mathbb{R}$ is a regular function. The probability density function and the characteristic function of X_2 are respectively given by:

$$p_{X_2}(x_2) \equiv \int_{\mathbb{R}} dx_1\ \delta(x_2 - h(x_1))\ p_{X_1}(x_1) \tag{40}$$

$$f_{X_2}(k) = \int_{\mathbb{R}} dx_1\ e^{ikh(x_1)}\ p_{X_1}(x_1) \tag{41}$$

Example (Scaling property of a normal distribution)**.** Let $X_1 = \mathbb{N}(\eta,\sigma^2)$ and $X_2 \equiv \alpha X_1$ with $\alpha\in\mathbb{R}$, then

$$\begin{aligned} f_{X_2}(k) &= \int_{\mathbb{R}} dx_1\ e^{i\alpha k x_1} p_{X_1}(x_1) \\ &= e^{ik\alpha\eta} e^{-\frac{\alpha^2\sigma^2k^2}{2}}. \end{aligned} \tag{42}$$

Hence we have

$$\alpha\ \mathbb{N}(\eta,\sigma^2) = \mathbb{N}(\alpha\eta,\alpha^2\sigma^2). \tag{43}$$

0.7 Two random variables

In this section we consider two random variables X_1, X_2 referring to the same stochastic experiment. This means that X_1 and X_2 are both defined with respect to the same probability space $(\Omega,\mathcal{F},\mathbb{P})$. The central object defining their property is the joint distribution.

Definition (Joint probability density function p_{X_1,X_2})**.** The joint probability density function of the random variables X_1, X_2 defined with respect to the probability space $(\Omega,\mathcal{F},\mathbb{P})$ is the function $p_{X_1,X_2}:\mathbb{R}^2\to\mathbb{R}$ such that

$$p_{X_1,X_2}(x_1,x_2)\ dx_1dx_2 = \mathbb{P}\left(\{x_1\le X_1\le x_1+dx_1\}\cap\{x_2\le X_2\le x_2+dx_2\}\right). \tag{44}$$

The single variable probability density functions are re-obtained through an operation called *reduction*.

Definition (Reduced density function p_X)**.** Given two random variables X_1, X_2 with joint probability density function p_{X_1,X_2}, the reduced probability density functions are defined as

$$p_{X_1}(x_1) \equiv \int_{\mathbb{R}} dx_2\ p_{X_1,X_2}(x_1,x_2), \tag{45}$$

$$p_{X_2}(x_2) \equiv \int_{\mathbb{R}} dx_1\ p_{X_1,X_2}(x_1,x_2) \tag{46}$$

In terms of characteristic functions one simply has

$$f_{X_1}(k_1) \equiv \int_{\mathbb{R}} dx_1\ e^{ik_1x_1}\ p_{X_1}(x_1) = f_{X_1,X_2}(k_1,0), \tag{47}$$

$$f_{X_2}(k_2) \equiv \int_{\mathbb{R}} dx_2\ e^{ik_2x_2}\ p_{X_2}(x_2) = f_{X_1,X_2}(0,k_2), \tag{48}$$

where

$$f_{X_1,X_2}(k_1,k_2) \equiv \int_{\mathbb{R}^2} dx_1dx_2\ e^{i(k_1x_1+k_2x_2)}\ p_{X_1,X_2}(x_1,x_2) \tag{49}$$

is the *joint characteristic function.*

Remark. In general, the joint probability density function p_{X_1,X_2} cannot be determined only from the knowledge of the reduced probability density function p_{X_1}, p_{X_2}.

Suppose the realization of a certain event $A \in \mathcal{F}$, with $X_1(A) = x_1$ (hence $\mathbb{P}(A) \neq 0$). It is then interesting to know if this fact influences or not the realization of another event B, with $X_2(B) = x_2$. A is called the *conditioning* event, while B is the *conditioned* event. The answer to this question is given by the conditional probability density function.

Definition (Conditional probability density function $p_{X_2|X_1}$)**.** Given two random variables X_1, X_2 with joint probability density function p_{X_1,X_2}, one could ask which is the probability of the event $\{x_2 \leq X_2 \leq x_2 + dx_2\}$ conditioned to the event $\{x_1 \leq X_1 \leq x_1 + dx_1\}$. According to what we said in the case of one random variable we can write

$$\mathbb{P}\{x_2 \leq X_2 \leq x_2 + dx_2 | x_1 \leq X_1 \leq x_1 + dx_1\} = \frac{\mathbb{P}\{x_2 \leq X_2 \leq x_2 + dx_2 \cap x_1 \leq X_1 \leq x_1 + dx_1\}}{\mathbb{P}\{x_1 \leq X_1 \leq x_1 + dx_1\}} \tag{50}$$

that, in terms of probability density functions becomes

$$p_{X_2|X_1}(x_2|x_1)dx_2 = \frac{p_{X_1,X_2}(x_2, x_1)dx_1dx_2}{p_{X_1}(x_1)dx_1} \tag{51}$$

giving the definition for $p_{X_2|X_1}(x_2|x_1)$:

$$p_{X_2|X_1}(x_2|x_1) \equiv \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_1}(x_1)}. \tag{52}$$

It it interesting to notice that given the conditional probability density function $p_{X_2|X_1}(x_2|x_1)$ and the reduced density function $p_{X_1}(x_1)$ it is possible to obtain the reduced density function for the variable X_2, namely

$$p_{X_2}(x_2) = \int_{\mathbb{R}} dx_1 \; p_{X_2|X_1}(x_2|x_1) \; p_{X_1}(x_1). \tag{53}$$

0.7.1 Moments of two random variables

Given two random variables X_1, X_2 with joint probability density function p_{X_1,X_2}, we have the following definitions:

Definition (Mixed moment of order $r_1 + r_2$, m_{r_1,r_2})**.** The mixed moment of order $r_1 + r_2 = 1, 2, \ldots$ is the real number:

$$m_{r_1,r_2} \equiv \mathbb{E}\{(X_1)^{r_1}(X_2)^{r_2}\} = \int_{\mathbb{R}^2} dx_1dx_2 \; x_1^{r_1} \; x_2^{r_2} \; p_{X_1,X_2}(x_1, x_2). \tag{54}$$

In particular we have $m_{1,0} = \eta_{X_1}$, $m_{0,1} = \eta_{X_2}$. $m_{1,1}$ is also called the *correlation*, C_{X_1,X_2}, between X_1 and X_2:

$$C_{X_1,X_2} \equiv \mathbb{E}\{X_1X_2\} = m_{1,1}. \tag{55}$$

As for the single variable case, the moments are easily calculated from the joint characteristic function:

$$m_{r_1,r_2} = \lim_{k_1,k_2 \to 0} (-i)^{r_1+r_2} \frac{\partial^{r_1+r_2} f_{X_1,X_2}(k_1, k_2)}{\partial k_1^{r_1} \partial k_2^{r_2}}. \tag{56}$$

Definition (Central mixed moment of order $r_1 + r_2$, μ_{r_1,r_2})**.** The central mixed moment of order $r_1 + r_2 = 1, 2, \ldots$ is the real number:

$$\mu_{r_1,r_2} \equiv \mathbb{E}\{(X_1 - \eta_{X_1})^{r_1}(X_2 - \eta_{X_2})^{r_2}\} = \int_{\mathbb{R}^2} dx_1dx_2 \; (x_1 - \eta_{X_1})^{r_1} \; (x_2 - \eta_{X_2})^{r_2} \; p_{X_1,X_2}(x_1, x_2). \tag{57}$$

In particular we have $\mu_{2,0} = \sigma^2_{X_1}$, $\mu_{0,2} = \sigma^2_{X_2}$. $\mu_{1,1}$ is also called the *covariance*, Cov_{X_1,X_2}, between X_1 and X_2:

$$Cov_{X_1,X_2} \equiv \mathbb{E}\{(X_1 - \eta_{X_1})(X_2 - \eta_{X_2})\} = \mu_{1,1}. \tag{58}$$

0.8 Conditional expectation

If (X_1, X_2) is a two-dimensional random variable with joint density probability function $p_{X_1X_2}(x_1, x_2)$, the expectation value of X_2 conditioned to $X_1 = x$ is given by

$$\mathbb{E}(\{X_2|x_1\}) = \int_{\mathbb{R}} x_2 p_{X_2|X_1}(x_2|x_1)dx_2 \tag{59}$$

Notice that $\mathbb{E}(\{X_2|x_1\})$ is a function of x. Similarly, the second central moment of X_2 given $X_1 = x_1$ is defined as

$$\mathbb{E}(\{(X_2 - \eta_{X_2})^2|x_1\}) \int_{\mathbb{R}} (x_2 - \eta_{X_2})^2 p_{X_2|X_1}(x_2|x_1)dx_2 \tag{60}$$

0.9 n random variables

The definition of the previous section generalizes straightforwardly to the case of $n = 2, 3, \ldots$ random variables. We have then the following self-explanatory formulas:

Definition (Joint probability density function $p_{X_1,\ldots X_n}$)**.**

$$\begin{gathered} p_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{R} \\ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\ dx_1 \cdots dx_n = \mathbb{P}(x_1 \leq X_1 \leq x_1 + dx_1 \cap \cdots \cap x_n \leq X_n \leq x_n + dx_n). \end{gathered} \tag{61}$$

Definition (Characteristic function of the joint probability density function $f_{X_1,\ldots X_n}$)**.**

$$\begin{gathered} f_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{C} \\ f_{X_1,\ldots X_n}(k_1, \ldots, k_n) \equiv \int_{\mathbb{R}^n} dx_1 \ldots dx_n\ e^{i(k_1x_1+\cdots+k_nx_n)}\ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \end{gathered} \tag{62}$$

Definition (Reduced density function $p_{X_1,\ldots,X_l}$)**.**

$$\boxed{p_{X_1,\ldots,X_l}(x_1, \ldots, x_l) \equiv \int_{\mathbb{R}^{n-l}} dx_{l+1} \cdots dx_n\ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \qquad l < n.} \tag{63}$$

Definition (Reduced characteristic function $f_{X_1,\ldots,X_l}$)**.**

$$f_{X_1,\ldots,X_l}(k_1, \ldots, k_l) = f_{X_1,\ldots,X_n}(k_1, \ldots, k_l, 0, \ldots, 0). \tag{64}$$

Definition (Conditional probability density function $p_{X_n,\ldots,X_{l+1}|X_l,\ldots,X_1}$)**.**

$$\boxed{p_{X_n,\ldots,X_{l+1}|X_l,\ldots,X_1}(\underbrace{x_n, \ldots, x_{l+1}}_{\text{left variables}} | \underbrace{x_l, \ldots, x_1}_{\text{right var.}}) \equiv \frac{p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{p_{X_1,\ldots,X_l}(x_1, \ldots, x_l)}.} \tag{65}$$

It is easy to prove the following important rules for eliminating a "left" or a "right" variable.

Remark (Elimination of "left" variables)**.**

$$\boxed{\int_{\mathbb{R}^2} dx_2dx_3\ p_{X_4,X_3,X_2|X_1}(x_4, x_3, x_2|x_1) = p_{X_4|X_1}(x_4|x_1).} \tag{66}$$

Remark (Elimination of "right" variables)**.**

$$\boxed{\int_{\mathbb{R}^2} dx_2dx_3\ p_{X_4|X_3,X_2,X_1}(x_4|x_3, x_2, x_1)\ p_{X_3,X_2|X_1}(x_3, x_2|x_1) = p_{X_4|X_1}(x_4|x_1).} \tag{67}$$

An important application of rule (67) is the integral Chapman-Kolmogorov formula.

Definition (Integral Chapman-Kolmogorov formula)**.**

$$\boxed{p_{X_3|X_1}(x_3|x_1) = \int_{\mathbb{R}} dx_2\; p_{X_3|X_2,X_1}(x_3|x_2,x_1)\; p_{X_2|X_1}(x_2|x_1).} \tag{68}$$

Exercise. Calculate the characteristic function of a zero-mean multivariate Gaussian distribution for the variables $\{X_i\}_i = 1, 2, \ldots, n$. Use the result to prove Wick's theorem, i.e., that all higher order moments can be expressed in terms of $\langle x_i^2 \rangle$ and $\langle x_i x_j \rangle$.

0.9.1 Function of many random variables

Given the random variables $X_1, \ldots, X_n$, consider the new random variable

$$X \equiv h(X_1, \ldots, X_n), \tag{69}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is a regular function. The probability density function and the characteristic function of X are respectively given by:

$$p_X(x) \equiv \int dx_1 dx_2 \cdots dx_n\; \delta(x - h(x_1, \ldots, x_n))\; p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \tag{70}$$

$$f_X(k) = \int dx_1 dx_2 \cdots dx_n\; e^{ikh(x_1,\ldots,x_n)}\; p_{X_1,\ldots,X_n}(x_1, \ldots, x_n). \tag{71}$$

Example (Sum of n random variables S_n)**.** Let $S_n \equiv X_1 + \cdots + X_n$, then its characteristic function is given by

$$f_{S_n}(k) = f_{X_1,\ldots,X_n}(k, k, \ldots, k). \tag{72}$$

0.9.2 Independent random variables

Definition (Independent events)**.** Two events $A, B \in \mathcal{F}$ are said independent when

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{73}$$

An analogous definition holds for random variables

Definition (Independent random variables)**.** n random variables $X_1, \ldots, X_n$ are said independent when

$$\mathbb{P}(X_1 \le x_1 \cap \cdots \cap X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) \tag{74}$$

for any $(x_1, \ldots, x_n)$.

For independent random variables we have then the following properties:

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n) \tag{75}$$

$$f_{X_1,\ldots,X_n}(k_1, \ldots, k_n) = f_{X_1}(k_1) \cdots f_{X_n}(k_n) \tag{76}$$

$$p_{X_n|X_{n-1},\ldots,X_1}(x_n|x_{n-1}, \ldots, x_1) = p_{X_n}(x_n). \tag{77}$$

Example (n independent Gaussian variables)**.** The probability density function of n independent identically distributed Gaussian random variables is given by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \equiv \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{x_1^2+\cdots+x_n^2}{2\sigma^2}} \tag{78}$$

Example (Stability property of a normal distribution)**.** Let $X_1 = \mathbb{N}(\eta_1, \sigma_1^2)$ and $X_2 = \mathbb{N}(\eta_2, \sigma_2^2)$ be two independent variables and $S_2 \equiv X_1 + X_2$, then

$$\begin{aligned} f_{S_2}(k) &= f_{X_1}(k) f_{X_2}(k) \\ &= e^{ik\eta_1} e^{-\frac{\sigma_1^2 k^2}{2}} e^{ik\eta_2} e^{-\frac{\sigma_2^2 k^2}{2}} && (79) \\ &= e^{ik(\eta_1+\eta_2)} e^{-\frac{(\sigma_1^2+\sigma_2^2)k^2}{2}}. && (80) \end{aligned}$$

Hence we have

$$\mathbb{N}(\eta_1, \sigma_1^2) + \mathbb{N}(\eta_2, \sigma_1^2) = \mathbb{N}(\eta_1 + \eta_2, \sigma_1^2 + \sigma_2^2). \tag{81}$$

0.10 Infinitely divisible Distributions

Definition. A random variable X is infinitely divisible if $\forall N \in \mathbb{N}$, it can represented by a sum

$$X = X_1 + X_2 + \cdots X_N \tag{82}$$

of i.d. random variables, $(X_1, \cdots, X_N)$

TODO

0.11 Central limit theorem

Consider the sum $S_n = X_1 + \cdots + X_n$ of independent random variables identically distributed $(p_{X_i}(x_i) = p_X(x)\ \forall i = 1, 2, \ldots n)$ with mean η and variance σ^2. Moreover let us assume that $\mathbb{E}\{X_i^2\} < \infty$. Then

$$p_{\frac{(S_n - n\eta)}{\sigma\sqrt{n}}} \underset{n\to\infty}{\sim} = \mathbb{N}(0, 1) \tag{83}$$

Sketch of the proof. Let us consider the set of random variables $\zeta_i = X_i - \eta$. Since the X_i are independent also the ζ_i are independent. Moreover $\mathbb{E}\{\zeta_i\} = 0$ and $Var\{\zeta_i\} = \sigma^2$. The characteristic function for the probability density $p_\zeta(x)$ is given by

$$\begin{aligned} f_\zeta(k) &= \sum_{n=0}^{\infty} \frac{1}{r!} \left.\frac{d^r f}{dk^r}\right|_{k=0} k^r \\ &= \sum_{n=0}^{\infty} \frac{(i)^r}{r!} m_r k^3 \\ &= 1 - \frac{m_2}{2} k^2 + \cdots = 1 - \frac{\sigma^2}{2} k^2 + \cdots && (84) \end{aligned}$$

since the first moment $\mathbb{E}\{\zeta_i\}$ is zero by construction. We now consider the random variable $S_n - \eta n = (X_1 - \eta) + (X_2 - \eta) + \cdots + (X_n - \eta)$. Because of independence we have

$$\begin{aligned} f_{S_n - n\eta}(k) &= [f_\zeta(k)]^n && (85) \\ &= \left[1 - \frac{(n\sigma^2 k^2)}{2n} + o\left(\frac{1}{n}\right)\right]^n && (86) \\ &\underset{n\to\infty}{\sim} \exp\left(-\frac{1}{2}(\sigma k\sqrt{n})^2\right). && (87) \end{aligned}$$

Hence

$$(S_n - n\eta) \in \mathbb{N}(0, n\sigma^2) \quad \text{for} \quad n \to \infty. \tag{88}$$

Now let us consider the function $h(S_n - n\eta) = \frac{S_n - n\eta}{\sigma\sqrt{n}} = \alpha(S_n - n\eta)$. From the scaling property of the normal distribution, $\alpha\mathbb{N}(\eta, \sigma^2) = \mathbb{N}(\alpha\eta, \alpha^2\sigma^2)$, we finally have

$$\frac{(S_n - n\eta)}{\sigma\sqrt{n}} \in \mathbb{N}(0, 1) \quad \text{for} \quad n \to \infty \qquad \text{(QED)}. \tag{89}$$

Remark (Stable laws and CLT)**.** The theorem above requires that the variance of X_i is finite. For random variables X_i with unbounded variances one can show the following: If there exists $\{a_n\}$ and $\{b_n\}$ such that

$$\mathbb{P}\left\{a_n(S_n - b_n) \leq x\right\} \to G(x) \qquad \text{as} \quad n \to \infty, \tag{90}$$

then the distribution $G(x)$ is stable. Clearly if the variance is finite than $G(x)$ is the Gaussian with $b_n = \eta = \mathbb{E}\{X_1\}$ and $a_n = \sqrt{Var\{X_1\}}$.

0.12 Statistics of Extrema

In many cases one is interested in estimating the maximum or minimum of a set of random variables. Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables and let,

$$M_n = \max\{X_1, X_2, \cdots, X_n\}. \tag{91}$$

The goal is to study the distribution of M_n as $n \to \infty$. Let us first look at some examples to have an idea.

Example. Assume that X_i is expnentially distributed, i.e., if we denote by $p(x)$ the PDF of X_i we have

$$p(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{92}$$

In terms of comulative distribution this means

$$P\{X_i < x\} = \int_{-\infty}^{x} e^{-y} dy = 1 - e^{-x}, \qquad \text{for } x > 0. \tag{93}$$

For any $x > 0$ we then have

$$\begin{aligned} P\{M_n \leq x\} &= P\{X_i \leq x \quad \forall i = 1, 2, \cdots, n\} \\ &= \prod_{i=1}^{n} P\{X_i < x\} = (1 - e^{-x})^n, \end{aligned} \tag{94}$$

where in the last equation we have used the independence of the $\{X_i\}_{i=1}^n$. The above equation remains true even if x depends on n. For example let us chose $x = x_n$ such that $(1 - e^{-x_n})^n$ has a non trivial limit. Let

$$x_n = -\log(e^{-x}) + \log n = x + \log n. \tag{95}$$

Hence

$$P\{M_n \leq x\} = (1 - e^{-x_n})^n = \left(1 - \frac{e^{-x}}{n}\right)^n \to e^{-e^{-x}}, \quad \text{as } n \to \infty. \tag{96}$$

In other words

$$P\{M_n \leq x + \log n\} \to e^{-e^{-x}}, \quad \text{as } n \to \infty. \tag{97}$$

The last equation tell us that M_n grows as $\log n$ as $n \to \infty$.

Example. Suppose that X_i is uniformly distributed on $[0, 1]$,i.e.

$$p(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{98}$$

For this distribution we expect that $M_n \to 1$ as $n \to \infty$. The question is: how does M_n converges to this limit ? Suppose we consider $x_n = 1 - x/n$, then

$$P\{M_n \leq x_n\} = \left(1 - \frac{x}{n}\right)^n \to e^{-x}, \quad \text{as } n \to \infty. \tag{99}$$

In other words

$$P\{n(M_n - 1) \leq x\} \to e^{-|x|} \qquad \text{for } x \leq 0. \tag{100}$$

This implies that $1 - M_n = O(1/n)$ as $n \to \infty$.

For the general case we have the following result

Theorem 0.12.1. If there exists {} and {} such that

$$P\{a_n(M_n - b_n) \le x\} \to G(x) \qquad \text{as } n \to \infty, \tag{101}$$

then $G(x)$ must be one of the following types:

Type I:

$$G(x) = e^{-e^{-x}}. \tag{102}$$

This distribution is known as the *Gumbel* distribution.

Type II:

$$G(x) = \begin{cases} 0 & \text{if } x \le 0 \\ e^{-x^{-\alpha}} & \text{if } x > 0 \end{cases} \tag{103}$$

for some $\alpha > 0$;

Type III:

$$G(x) = \begin{cases} e^{-|x|^{\alpha}} & \text{if } x \le 0 \\ 1 & \text{if } x > 0 \end{cases} \tag{104}$$

for some $\alpha > 0$.

In the examples above we have $b_n = \log n$, $a_n = 1$ for the exponential distribution which is a type I situation, and $b_n = 1$, $a_n = n$ for the uniform distribution which is a type III situation with $\alpha = 1$.

Exercise. Verify that

$$a_n = \sqrt{2 \log n}, \qquad b_n = \sqrt{2 \log n}(\log \log n + \log 4\pi) \tag{105}$$

for the normal distribution $\mathbb{N}(0, 1)$.

Problems

1. Show that if the zero-mean random variable ξ is Gaussian-distributed, the kurtosis (or flatness) defined as

$$\frac{\mathbb{E}\{\xi^4\}}{\mathbb{E}\{\xi^2\}^2}$$

equals 3.

2. Let $y = f\cos\theta$, where f and θ are independent zero-mean random variables, with f chosen from a Gaussian distribution of unit variance and θ from the uniform distribution on $[-\pi, \pi]$. Prove that y is not a Gaussian variable.

3. Let $z = f\cos\theta + g\sin\theta$ with f and θ as above, and g chosen from a Gaussian distribution of unit variance, independent of f and θ. Is z Gaussian ?

4. Find the distribution of $\cos\theta$ for θ uniformly distributed in $[0, 2\pi]$.

5. Two boxes of volumes V_1 and V_2 contain N non-interacting molecules and are connected through a hole. Calculate the probability of finding N_1 molecules in V_1.

6. Show that

- the characteristic function of a normal distribution $\mathbb{N}(\eta, \sigma^2)$ is

$$f_\xi(k) = e^{ik\eta}e^{-\frac{\sigma^2 k^2}{2}}, \quad k \in \mathbb{R}, \tag{106}$$

- the one of a Cauchy distribution is

$$f_\xi(k) = e^{ikx_0}e^{-\alpha|k|}, \quad k \in \mathbb{R}, \tag{107}$$

- the one of a uniform distribution is

$$f_\xi(k) = \frac{e^{ikx_2} - e^{ikx_1}}{ik(x_2 - x_1)}, \quad k \in \mathbb{R}. \tag{108}$$

- whereas for a binomial distribution we have

$$f_\xi(k) = \left(q + pe^{ik}\right)^n, \quad k \in \mathbb{R}. \tag{109}$$

7. A binary source generates 1's and 0's randomly, respectively with probablity 0.6 and 0.4.

(A) Which is the probability that two 1's and three 0's occur in a five bits sequence ?

(B) Which is the probability that at least three 1's occur in a five bits sequence ?

8. A fair coin is tossed ten times. Compute the probability of having 5 or 6 head in this experiment.

9. The probability density function of a tw0-dimensional random variable (ξ_1, ξ_2) si given by

$$p_{\xi_1\xi_2}(x_1, x_2) = \begin{cases} kx_1x_2 & 0 < x_1 < 1, \quad 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{cases} \tag{110}$$

where k is a constant.

(A) Compute the value of k

(B) Are ξ_1 and ξ_2 independent r.v. ?

(C) Compute the probability $\mathbb{P}\{\xi_1 + \xi_2 < 1\}$.

10. Let (ξ_1, ξ_2, ξ_3) be a three-dimensional random variable with probability density

$$f_{\xi_1,\xi_2\xi_3}(x_1, x_2, x_3) = \begin{cases} ke^{-(ax_1+bx_2+cx_3)} & x_1 > 0, \quad x_2 > 0, \quad x_3 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{111}$$

where a, b, c are constants.

(A) Compute the value of k.

(B) Compute the jont marginal distribution for ξ_1 and ξ_2.

(C) Compute the marginal distribution for ξ_1.

(D) Are ξ_1,ξ_2 and ξ_3 independent ?

11. Two independent stochastic variables X and Y have normal even distributions with variance σ_x^2 and σ_y^2 respectively. Find the joint probability density functions for the stochastic variables $V = X + Y$ and $W = X - Y$.

Bibliography

[1] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, (Clarendon Press, Oxford 1982).

[2] R. Durrett, *Probability: Theory and Examples.*

[3] J. Jacood and Protter, *Probability Essentials.*

[4] D. Williams, *Probability with Martingales.*

[5] Billingsley, *Probability and Measure.*

[6] J. Pitman, *Probability.*

[7] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, (McGraw-Hill, Singapore 1984).